Adaptacyjne i nieortogonalne systemy dostępu wielokrotne w 6G

Jednym z trendów technologicznych w komunikacji bezprzewodowej jest to, że szerokość pasma sygnału staje się coraz szersza, co ma na celu obsługę wyższej szybkości transmisji. Jednak wraz ze zmniejszaniem się okresu symbolu, rozproszenie opóźnienia w kanale zanikania wielościeżkowego powoduje poważne zakłócenia między symbolami i znacznie ogranicza osiągalną szybkość transmisji. W tradycyjnej transmisji pojedynczej nośnej korektor z kilkoma setkami odczepów może być konieczny do skutecznego złagodzenia zakłóceń między symbolami w systemie szerokopasmowym, który jest zbyt złożony, aby wdrożyć go w praktycznym systemie. W tym względzie modulacja wielonośnikowa zapewnia wydajną alternatywę poprzez podzielenie sygnału szerokopasmowego na zestaw ortogonalnych sygnałów wąskopasmowych. Ze względu na zdolność radzenia sobie z wielodrożnym zanikiem częstotliwości bez potrzeby złożonej korekcji i prostej implementacji poprzez wykorzystanie cyfrowej transformaty Fouriera, ortogonalne multipleksowanie z podziałem częstotliwości (OFDM), jako rodzaj modulacji wielonośnej, stało się najbardziej dominującą techniką projektowania przebiegów dla przewodowych i bezprzewodowych systemów komunikacyjnych w ciągu ostatnich dwóch dekad od czasu jej pierwszej propozycji przez Changa . W rezultacie zostało ono szeroko zastosowane w wielu znanych standardach, np. Digital Subscriber Line (DSL), Digital Video Broadcasting-Terrestrial (DVB-T), Wi-Fi, WiMAX, LTE, LTE-Advanced i 5G NR. Z drugiej strony sieć komórkowa musi pomieścić wielu aktywnych abonentów jednocześnie w skończonej ilości zasobów czasowo-częstotliwościowych. Dlatego też efektywne przydzielanie zasobów radiowych pomiędzy użytkownikami jest krytycznym aspektem projektowania zarówno kanałów łącza w górę, jak i w dół, ponieważ przepustowość jest zwykle ograniczona i droga. Udział kanału komunikacyjnego wśród wielu użytkowników rozproszonych geograficznie jest określany jako wielokrotny dostęp. Konwencjonalne techniki wielokrotnego dostępu ortogonalnie dzielą wymiary sygnalizacji w dziedzinie czasu, dziedzinie częstotliwości lub dziedzinie kodu. Jako rozszerzenie transmisji OFDM w celu wdrożenia systemu wielodostępnego, ortogonalny dostęp z podziałem częstotliwości (OFDMA) zapewnia wydajną i elastyczną technologię wielokrotnego dostępu poprzez jednoczesne wykorzystanie zasobów czasowo-częstotliwościowych. Historycznie większość systemów mobilnych opierała się na ortogonalnych technikach wielokrotnego dostępu w celu prostego projektowania systemów i implementacji odbiorników o niskiej złożoności. Aby sprostać heterogenicznym wymaganiom dotyczącym masowej łączności, wysokiej wydajności widmowej, niskiego opóźnienia i zwiększonej uczciwości, system 5G zaakceptował nową technikę zwaną nieortogonalnym dostępem wielokrotnym (NOMA) jako jedną ze swoich metod wielokrotnego dostępu. W przeciwieństwie do konwencjonalnych schematów ortogonalnych, kluczową cechą wyróżniającą NOMA jest obsługa większej liczby użytkowników niż liczba jednostek zasobów ortogonalnych przy pomocy nieortogonalnego współdzielenia zasobów, za cenę wyrafinowanej eliminacji zakłóceń między użytkownikami w odbiorniku. Przewiduje się, że zarówno ortogonalne, jak i NOMA będą dalej rozwijane i odegrają kluczową rolę w nadchodzącym projekcie systemu 6G. W tym rozdziale zostaną przedstawione techniki ortogonalne i NOMA, składające się z

● Modelowania selektywnego częstotliwościowo kanału zanikającego w bezprzewodowej komunikacji szerokopasmowej.

● Podstawy modulacji wielonośnej, w tym filtry syntezy i analizy, implementacja wielofazowa i filtr wielonośnikowy banku.

● Kompleksowe wprowadzenie do techniki OFDM, składającej się z jej podstawowej zasady, wydajnej implementacji poprzez cyfrową transformację Fouriera, wstawianie cyklicznego prefiksu, przetwarzanie sygnału w dziedzinie częstotliwości i tłumienie emisji poza pasmem.

● Transmisja OFDMA w łączu w dół i jednonośnikowa FDMA w łączu w górę.

● Podstawowa zasada i zaleta zróżnicowania opóźnień cyklicznych.

● OFDMA wielokomórkowe i bezkomórkowe masywne MIMO-OFDMA.

● Podstawy NOMA, w tym zasady NOMA w domenie mocy i NOMA w domenie kodu, transmisja superpozycji wielu użytkowników i transmisja bez przyznawania.

Kanał zanikania selektywny częstotliwościowo

Kanał zanikania wielodrogowego można opisać za pomocą odpowiedzi w czasie t na impuls wejściowy w czasie t − τ, mianowicie

$$h(\tau, t) = \sum_{l=1}^{L} a_l(t)\delta\left(\tau - \tau_l(t)\right), \quad (1)$$

gdzie $a_l(t)$ i $\tau_l(t)$ oznaczają odpowiednio tłumienie i opóźnienie propagacji l-tej ścieżki w czasie t, a L jest całkowitą liczbą możliwych do rozwiązania ścieżek. To wyrażenie jest całkiem ładne. Oznacza to, że wpływ użytkowników mobilnych, dowolnie przesuwających reflektory i absorbery oraz wszystkie zawiłości rozwiązywania równań Maxwella ostatecznie sprowadzają się do relacji wejście–wyjście, która jest reprezentowana jako odpowiedź impulsowa liniowego filtra kanałowego zmieniającego się w czasie. W szczególnej sytuacji, gdy nadajnik, odbiornik i otoczenie są nieruchome, opóźnienia tłumienia i propagacji nie zmieniają się w czasie i mamy liniowy kanał niezmienny w czasie z odpowiedzią impulsową

$$h(\tau) = \sum_{l=1}^{L} a_l\delta\left(\tau - \tau_l\right) \quad (2)$$

Praktyczna komunikacja bezprzewodowa to transmisja w paśmie przepustowym, która jest realizowana w paśmie o częstotliwości nośnej f_c. Jednak większość przetwarzania sygnału w komunikacji bezprzewodowej, takiego jak kodowanie kanału, modulacja, wykrywanie, synchronizacja i szacowanie, jest zwykle implementowana w paśmie podstawowym. Stąd sensowne jest uzyskanie złożonego modelu równoważnego pasma podstawowego

$$h_b(\tau, t) = \sum_{l=1}^{L} a_l(t)\delta\left(\tau - \tau_l(t)\right) e^{-2\pi j f_c \tau_l(t)}. \quad (3)$$

Następnym krokiem jest konwersja kanału o czasie ciągłym na kanał o czasie dyskretnym. Postępując zgodnie z twierdzeniem o próbkowaniu, możemy utworzyć bardziej użyteczny model kanału o czasie dyskretnym, ustalając ζ-ty odczep filtra kanału w (dyskretnym) czasie n, tj.

$$h_\zeta[n] = \sum_{l=1}^{L} a_l(nT_s)e^{-2\pi j f_c \tau_l(nT_s)}\mathrm{sinc}\left(\zeta - \frac{\tau_l(nT_s)}{T_s}\right), \quad \zeta = 0, 1, \ldots, Z-1 \quad (4)$$

gdzie $T_s = 1/B_w$ oznacza okres próbkowania z szerokością pasma przesyłanego sygnału B_w, a funkcja sinc jest zdefiniowana jako

$$\mathrm{sinc}(t) := \frac{\sin(\pi t)}{\pi t} \quad (5)$$

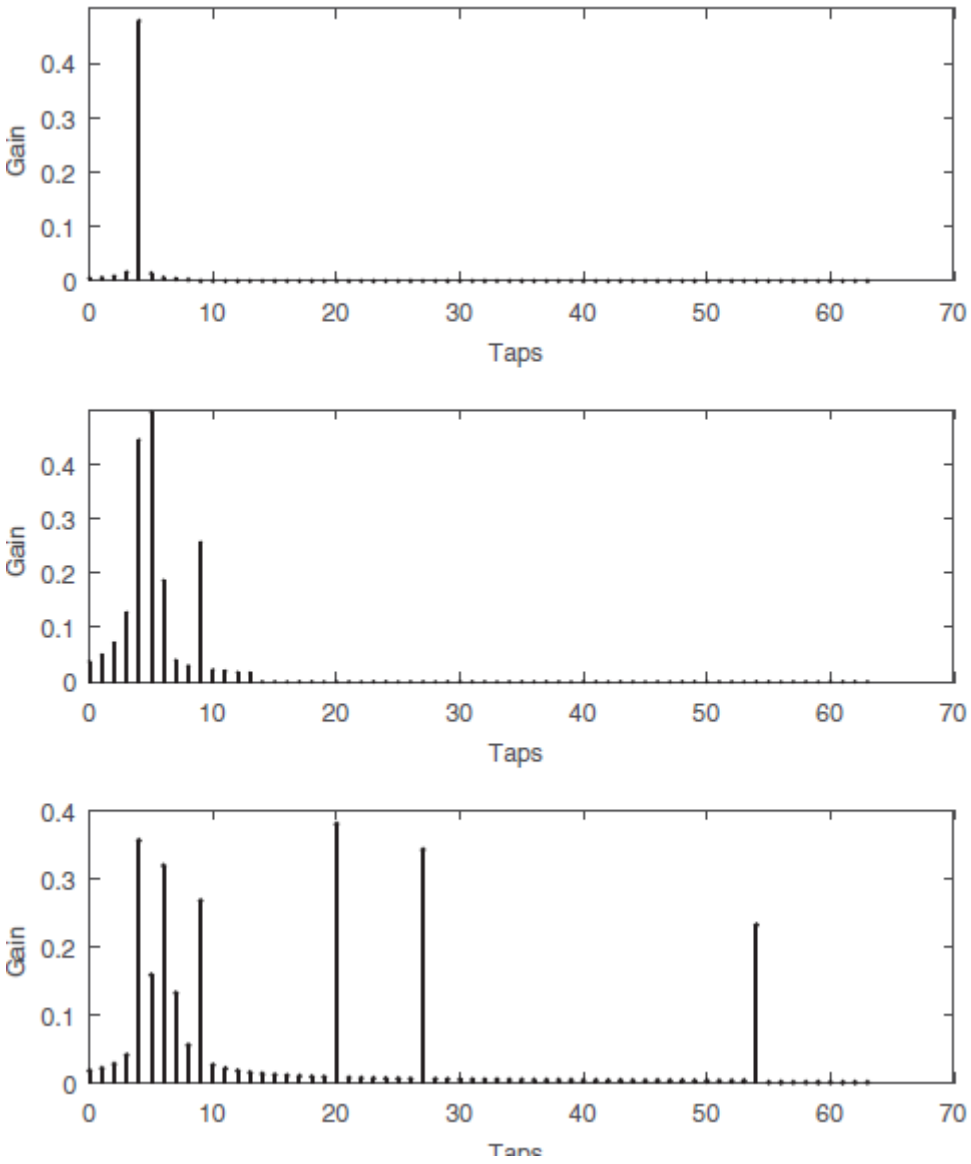
W szczególnym przypadku, gdy wzmocnienia i opóźnienia ścieżek są niezmienne w czasie, równanie (4) można uprościć do

$$h_\zeta = \sum_{l=1}^{L} a_l e^{-2\pi j f_c \tau_l} \mathrm{sinc}\left(\zeta - \frac{\tau_l}{T_s}\right) \quad (6)$$

Następnie zależność wejścia i wyjścia w czasie dyskretnym równoważnego systemu pasma podstawowego można wyrazić wzorem

$$y[n] = \sum_{\zeta=0}^{Z-1} h_\zeta[n] x[n-\zeta], \quad (7)$$

zakładając, że filtr kanałowy ma nieskończoną długość Z, która jest określana przez rozproszenie opóźnienia i częstotliwość próbkowania systemu. Rysunek ilustruje kilka przykładów odgałęzień filtra wygenerowanych z modelu kanału 3GPP Extended Typical Urban (ETU) przy różnych częstotliwościach próbkowania.



Przy wyższej częstotliwości próbkowania rozdzielczość staje się lepsza, co skutkuje dokładniejszym filtrem kanałowym opisanym przez więcej odgałęzień. Gdy impuls sygnału przechodzi przez kanał wielodrogowy, odebrany sygnał pojawi się jako ciąg impulsów, przy czym każdy impuls odpowiada ścieżce bezpośredniej lub ścieżce NLOS. Ważną cechą propagacji radiowej jest rozproszenie opóźnienia wielodrogowego lub dyspersja czasowa podniesiona z odrębnego czasu przybycia różnych ścieżek propagacji. Zakładając, że τ_1 w modelu kanału wielodrogowego podanym przez równanie (3) oznacza czas propagacji pierwszego przybywającego składnika wielodrogowego, minimalne opóźnienie nadmiarowe jest równe τ_1, jako opóźnienie odniesienia równe zero. Tymczasem czas propagacji ostatniego przybywającego składnika wielodrogowego wynosi τL. Rozprzestrzenianie się opóźnień można łatwo zmierzyć za pomocą różnicy czasu przybycia między najkrótszą i najdłuższą możliwą do rozwiązania ścieżką, zwanej również maksymalnym opóźnieniem nadmiarowym, w kategoriach

$$T_d := \tau_L - \tau_1 \quad (8)$$

Odpowiedź impulsowa kanału bezprzewodowego zmienia się zarówno w czasie, jak i częstotliwości, a rozproszenie opóźnienia określa, jak szybko zmienia się częstotliwość. Istnieje różnica faz 2π f ($\tau_i - \tau_j$) między dwoma dowolnymi składowymi wielodrogowymi i i j. Biorąc pod uwagę maksymalną różnicę faz między wszystkimi ścieżkami jako 2π fT_d, wielkość ogólnej odpowiedzi częstotliwościowej zmienia się znacząco, gdy różnica faz wzrasta lub maleje o wartość π. Tak więc szerokość pasma koherencji, która wskazuje szybkość zanikania kanału bezprzewodowego w dziedzinie częstotliwości, jest definiowana jako

$$B_c := \frac{1}{2T_d}. \quad (9)$$

W transmisji wąskopasmowej szerokość pasma przesyłanego sygnału jest zwykle znacznie mniejsza niż szerokość pasma koherencji, tj. $B_w \ll B_c$. Zatem zanikanie w całym paśmie jest silnie skorelowane, nazywane zanikaniem płaskim częstotliwościowo. W tym przypadku rozproszenie opóźnienia jest znacznie mniejsze niż okres symbolu T_s, a zatem pojedyncze dotknięcie wystarcza do przedstawienia filtru kanału, np.

$$h = -1.3162 + 0.3671i \quad (10)$$

jak pokazano na najwyższym wykresie na rysunku . W związku z tym relacja wejścia–wyjścia w równaniu (10.7) jest uproszczona do

$$y[n] = h[n]x[n]. \quad (11)$$

Przeciwnie, jeśli szerokość pasma sygnału $B_w \gg B_c$, dwa punkty częstotliwości oddzielone o więcej niż szerokość pasma koherencji wykazują mniej więcej niezależną odpowiedź. Tak więc komunikacja szerokopasmowa cierpi na zanik częstotliwości selektywny i interferencję między symbolami (ISI). Opóźnienie wielościeżkowe rozprzestrzenia się na wiele symboli, a filtr kanału może być reprezentowany tylko przez serię odczepów, np.

$$\mathbf{h} = [-1.316 + 0.367i, -0.144 - 0.08161i, 0.0772 + 0.0243i, -0.0515 - 0.014i$$
$$0.0386 + 0.0097i, -0.0308 - 0.0074i, 0.0257 + 0.0060i, -0.0220 - 0.0051i$$
$$0.0192 + 0.0044i, -0.0171 - 0.0038i, 0.0154 + 0.0034i, -0.0140 - 0.0031i$$
$$0.0128 + 0.0028i, -0.0118 - 0.0026i, 0.0110 + 0.0024i, -0.0102 - 0.0022i] \quad (12)$$

W komunikacji bezprzewodowej mechanizmy łagodzenia ISI odgrywają kluczową rolę w projektowaniu formatowania sygnału szerokopasmowego i struktury odbiornika. Można zastosować kilka technik w celu złagodzenia zniekształceń spowodowanych rozproszeniem opóźnienia wielodrożnego, w tym wyrównanie pojedynczej nośnej, rozproszone widmo i modulację wielonośnikową. Pierwsze dwie techniki są klasyczne i można je znaleźć w literaturze, takiej jak [Goldsmith, 2005]. Poniższa sekcja pokrótce przedstawi zasadę modulacji wielonośnej, mając na celu dostarczenie czytelnikom samodzielnej ilustracji.

Modulacja wielonośnikowa

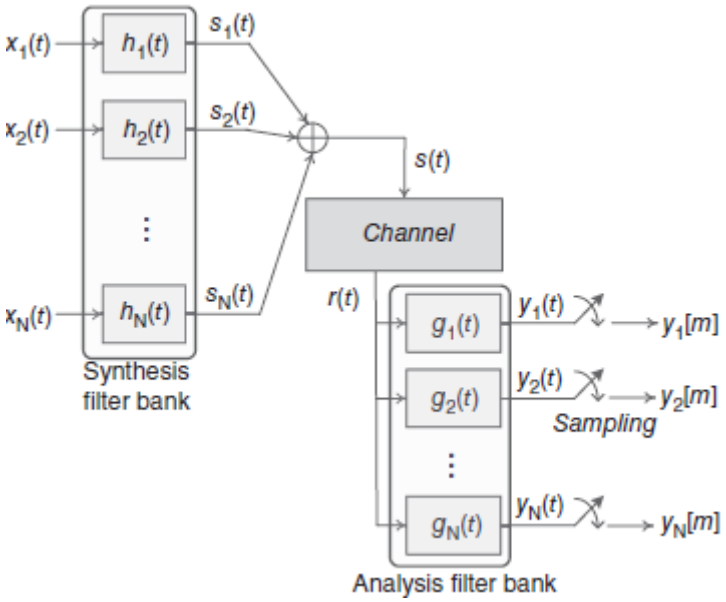
Jednym z trendów technologicznych w komunikacji bezprzewodowej jest to, że szerokość pasma sygnału staje się coraz szersza, aby obsługiwać wyższą szybkość transmisji]. W konwencjonalnej transmisji pojedynczej nośnej wyższa szerokość pasma w dziedzinie częstotliwości odpowiada krótszemu okresowi symbolu w dziedzinie czasu. Wraz ze zmniejszaniem okresu symbolu rozproszenie opóźnienia w kanale zanikania wielodrożnego zwiększa poważne ISI i znacznie ogranicza osiągalną szybkość transmisji. Tradycyjnie filtr cyfrowy, zwany korektorem, jest stosowany w odbiorniku w celu odwrócenia zniekształceń występujących w kanale. Liczba odczepów filtra wymaganych dla korektora jest proporcjonalna do szerokości pasma sygnału. Korektor z kilkoma setkami odczepów może być konieczny do skutecznego złagodzenia ISI w sygnale o ekstremalnie dużej szerokości pasma. Jest to zbyt skomplikowane, aby wdrożyć to w praktycznym systemie. Dlatego społeczność bezprzewodowa musi znaleźć alternatywę, która zastąpi transmisję pojedynczej nośnej. Modulacja wielonośnikowa (MCM) to szerokopasmowa technika komunikacyjna, w której sygnał szerokopasmowy jest dzielony na zestaw ortogonalnych sygnałów wąskopasmowych. Okres symbolu sygnału wąskopasmowego jest znacznie wydłużony i jest znacznie dłuższy niż sygnał szerokopasmowy. W ten sposób efekt ISI może zostać złagodzony w systemie MCM, jeśli rozproszenie opóźnienia stanie się nieistotne w porównaniu z wydłużonym okresem symbolu.

Filtry syntezy i analizy

W modulacji wielonośnej, tablica filtrów, znana również jako bank filtrów, jest stosowana w celu syntezy sygnałów wielonośnych w nadajniku. W związku z tym inny bank filtrów jest używany do analizy odebranych sygnałów wielonośnych w odbiorniku. Gdy sygnał x(t) przechodzi przez kanał z odpowiedzią impulsową h(t), wynikowy sygnał jest podany przez

$$s(t) = h(t) * x(t), \quad (13)$$

gdzie $*$ oznacza splot liniowy. Jak pokazano na rysunku ,



bank filtrów syntezy składa się z tablicy filtrów oznaczonych jako $h_n(t)$, n = 1, 2,…,N. Każdy filtr niezależnie odpowiada na swój impuls, zgodnie z

$$s_n(t) = h_n(t) * x_n(t). \quad (14)$$

Ich sygnały wyjściowe są sumowane, syntetyzując sygnał złożony, który jest podawany przez

$$s(t) = \sum_{n=1}^{N} s_n(t) = \sum_{n=1}^{N} h_n(t) * x_n(t). \quad (15)$$

Chociaż $h_n(t)$ może być w zasadzie dowolnym możliwym filtrowaniem, odpowiedź impulsowa filtru syntezy jest specjalnie zaprojektowana do przetwarzania sygnału wejściowego dla każdej podnośnej w systemie MCM. Transmitowany sygnał na n-tej podnośnej jest wyrażony przez

$$x_n(t) = \sum_{m=-\infty}^{\infty} u_{m,n}\delta(t - mT), \quad n = 1, \ldots, N, \quad (16)$$

gdzie $u_{m,n}$ oznacza symbol niosący informacje odpowiadający jednostce zasobu czasowo-częstotliwościowego nad n-tą podnośną w trakcie m-tego okresu symbolu, N oznacza całkowitą liczbę podnośnych, a T jest okresem symbolu. Podstawienie równania (16) do równania (15) daje ciągły sygnał pasma podstawowego o wielu nośnych

$$s(t) = \sum_{m=-\infty}^{+\infty} \sum_{n=1}^{N} u_{m,n} h_n(t - mT). \quad (17)$$

Aby osiągnąć ortogonalność między podnośnymi, odstęp między podnośnymi oznaczony jako Δf musi być całkowitą wielokrotnością odwrotności okresu symbolu. Ogólnie rzecz biorąc, odstęp jest ustawiony na Δf = 1⁄T, aby zmaksymalizować wydajność widmową [Jiang i Kaiser, 2016]. Bez utraty ogólności częstotliwości podnośnych w równoważnym modelu pasma podstawowego można zapisać jako

$$f_n = n \triangle f = \frac{n}{T}, \quad n = 1, 2, \ldots, N \tag{18}$$

Filtry syntezy bazują na specjalnie zaprojektowanym prototypie filtra $p_T(t)$ i są modulowane przez częstotliwości podnośnych f_n w następujący sposób:

$$h_n(t) = p_T(t) e^{2\pi j f_n t + j\phi_n} = p_T(t) e^{2\pi j n \Delta f t + j\phi_n}, \tag{19}$$

gdzie ϕ_n oznacza przesunięcie fazowe. Podstawiając równanie (19) do równania (17), sygnał wielonośnikowy pasma podstawowego można zapisać jako

$$\begin{aligned} s(t) &= \sum_{m=-\infty}^{\infty} \sum_{n=1}^{N} u_{m,n} p_T(t - mT) e^{2\pi j n \Delta f (t - mT) + j\phi_n} \\ &= \sum_{m=-\infty}^{\infty} \sum_{n=1}^{N} u_{m,n} p_T(t - mT) e^{2\pi j n \Delta f t + j\phi_n} e^{-2\pi j n m \Delta f T} \\ &= \sum_{m=-\infty}^{\infty} \sum_{n=1}^{N} u_{m,n} p_T(t - mT) e^{2\pi j n \Delta f t + j\phi_n}, \end{aligned} \tag{20}$$

stosując

$$e^{-2\pi j n m \Delta f T} = e^{-2\pi j n m} = e^{j0} = 1. \tag{21}$$

Odpowiednio odbiornik jest wyposażony w bank filtrów analizy składający się z tablicy filtrów, które mają wspólny przychodzący sygnał wielonośnikowy r(t). Chociaż można przeprowadzić dowolne możliwe filtrowanie, każdy filtr analizy przetwarza inną podnośną odebranego sygnału r(t) w komunikacji wielonośnej. Pomijając dla uproszczenia osłabienia kanału i szum addytywny, sygnał wejściowy dla banku filtrów analizy jest równy wygenerowanemu sygnałowi banku filtrów syntezy, tj. r(t) = s(t). Podobnie jak filtry syntezy, filtry analizy są oparte na specjalnie zaprojektowanym prototypowym filtrze $p_R(t)$. Podobnie jak w równaniu (19), odpowiedź impulsową typowego filtra analizy można wyrazić za pomocą

$$g_k(t) = p_R(t) e^{-(2\pi j f_k t + j\phi_k)} = p_R(t) e^{-(2\pi j k \Delta f t + j\phi_k)}, \quad k = 1, 2, \ldots, N. \tag{22}$$

Wprowadzając r(t) do typowego filtra analizy $g_k(t)$, wynikowy sygnał jest obliczany przez

$$\begin{aligned} y_k(t) &= g_k(t) * r(t) \\ &= \sum_{m=-\infty}^{\infty} \sum_{n=1}^{N} u_{m,n} p_R(t) * p_T(t - mT) e^{2\pi j n \Delta f t + j\phi_n} e^{-2\pi j k \Delta f t - j\phi_k} \\ &= \sum_{m=-\infty}^{\infty} \sum_{n=1}^{N} u_{m,n} p_R(t) * p_T(t - mT) e^{2\pi j (n-k) \Delta f t + j(\phi_n - \phi_k)}. \end{aligned} \tag{23}$$

Aby prawidłowo odzyskać symbol niosący informację w każdej jednostce zasobu czasowo-częstotliwościowego, muszą być spełnione dwa główne kryteria:

- Brak zakłóceń międzynośnych (ICI) w dziedzinie częstotliwości
- Brak zakłóceń ISI w dziedzinie czasu

Po pierwsze, podnośne muszą stanowić ortogonalny zestaw bazowy w okresie symbolu, aby uniknąć generowania ICI, tj.

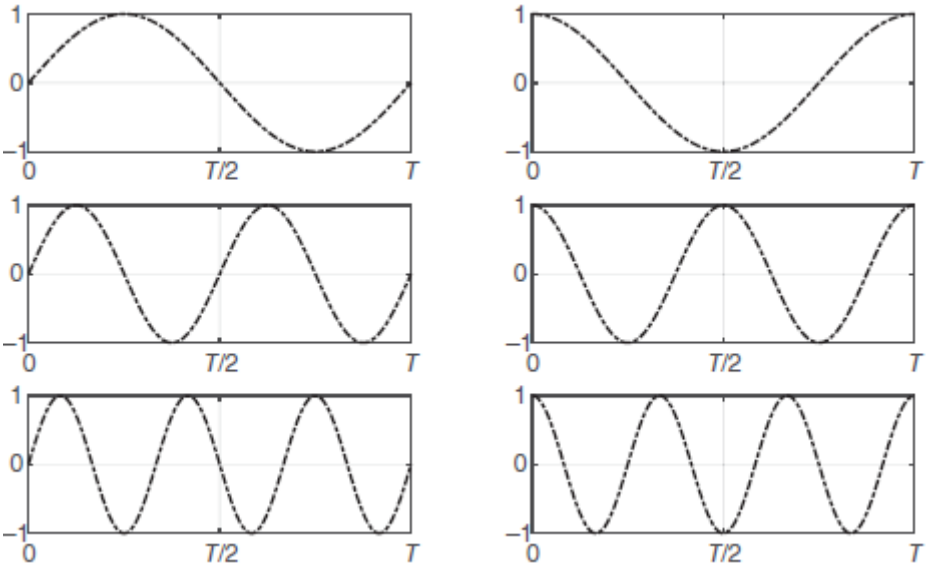
$$\frac{1}{T}\int_0^T e^{2\pi j(n-k)\Delta ft + j(\phi_n - \phi_k)}dt = \delta[n-k]. \quad (24)$$

Aby osiągnąć ortogonalność, odstęp między podnośnymi musi być równy całkowitej wielokrotności odwrotności okresu symbolu. Jak wspomniano wcześniej, odstęp jest zwykle wybierany jako $\Delta f = 1/T$ w celu maksymalizacji wydajności widmowej. Podnośne są oznaczane w formie wykładniczej $e^{2\pi jn\Delta ft}$, $t \in [0, T)$, z których każda ma dwie gałęzie: składowe w fazie (I) i kwadraturowe (Q). Symbol niosący informacje $u_{m,n}$ jest wartością zespoloną, tj. $u_{m,n} = a_{m,n} + jb_{m,n}$, gdzie $a_{m,n}$ i $b_{m,n}$ są liczbami rzeczywistymi. Zmodulowany sygnał na typowej podnośnej n jest oznaczany przez $\Re[u_{m,n}e^{2\pi jn\Delta ft}]$, co można przekształcić do postaci gałęzi I i Q jako:

$$a_{m,n}\cos(2\pi n\,\Delta\, ft) - b_{m,n}\sin(2\pi n\,\Delta\, ft). \quad (25)$$

Oznacza to, że rzeczywista część symbolu niosącego informację jest modulowana na sygnale gałęzi I podnośnej, podczas gdy część urojona jest przenoszona przez sygnał gałęzi Q. Jak pokazano na rysunku

,



fale sinusoidalne $\cos(2\pi n\Delta ft)$ i $\sin(2\pi n\Delta ft)$, n = 1, 2, 3 są wzajemnie ortogonalne w ciągu jednego okresu symbolu T, spełniając wymóg równania (24). Po drugie, dwa prototypowe filtry powinny spełniać warunek, że sygnał wyjściowy nie powoduje ISI w dziedzinie czasu. Nie oznacza to koniecznie, że nie ma żadnego nakładania się między kolejnymi symbolami, ale wymaga przynajmniej braku zakłóceń w punktach próbkowania, mianowicie

$$p_T(t) * p_R(t)\Big|_{t_s=iT} = \begin{cases} 1, & i = 0 \\ 0, & i \neq 0 \end{cases},$$

gdzie $t_s = iT$ oznacza punkty próbkowania na osi czasu. Warunek ten można spełnić, na przykład, stosując funkcję sinc w kształtowaniu impulsu.

Implementacja wielofazowa

Sygnał wielonośnikowy zawiera N podnośnych z odstępem między podnośnymi Δf , co daje szerokość pasma sygnału B_w = N Δ f . Zgodnie z twierdzeniem o próbkowaniu , interwał próbkowania T_s może być równy odwrotności szerokości pasma sygnału, tj.

$$T_s = \frac{1}{B_w} = \frac{1}{N \Delta f} = \frac{T}{N}. \quad (26)$$

Długość okresu symbolu wynosi T, więc odpowiadająca jej liczba próbek sygnału w każdym okresie symbolu wynosi T/T_s = N. Filtr prototypowy o dyskretnym czasie można uzyskać, próbkując filtr prototypowy o ciągłym czasie $p_T(t)$ z częstotliwością próbkowania T_s, co skutkuje

$$p_T[l] = p_T(lT_s), \quad l = 0, 1, \ldots, \mathfrak{L} - 1. \quad (27)$$

Możliwe jest, że długość prototypowego filtra w czasie dyskretnym jest większa niż okres symbolu i $\mathfrak{L}$ można wybrać jako całkowitą wielokrotność N. Zakładając, że współczynnik nakładania się wynosi K, co oznacza, że długość prototypowego filtra jest K razy dłuższa od okresu symbolu, tj.

$$\mathfrak{L} = KN. \quad (28)$$

W dziedzinie przetwarzania sygnałów transformacja Z jest ważnym narzędziem analitycznym, które przekształca sekwencję dyskretną w czasie w zespoloną reprezentację w dziedzinie częstotliwości. Transformację Z $p_T[l]$ oblicza się za pomocą

$$P_T(z) = \sum_{l=0}^{\mathfrak{L}-1} p_T[l] z^{-l} = \sum_{l=0}^{KN-1} p_T[l] z^{-l} \quad (29)$$

Zakładając l = k'N + n', gdzie k' = 0, 1,…, (K − 1) i n' = 0, 1,…, (N − 1), Równanie (29) można dalej przekształcić na

$$\begin{aligned} P_T(z) &= \sum_{n'=0}^{N-1} \sum_{k'=0}^{K-1} p_T\left[k'N + n'\right] z^{-(k'N+n')} \\ &= \sum_{n'=0}^{N-1} z^{-n'} \sum_{k'=0}^{K-1} p_T\left[k'N + n'\right] z^{-k'N}. \end{aligned} \quad (30)$$

Stanowiąc serię N podciągów $p^{n'}{}_T[k']$= k'N + n', n' = 0, 1,…, (N − 1), gdzie $p^{n'}{}_T[k']$ ma długość K i jest nazywany (n')-tą składową wielofazową prototypowego filtra $p_T[l]$. Transformata Z $p^{n'}{}_T[k']$, mianowicie

$$P^{n'}_T(z^N) = \sum_{k'=0}^{K-1} p_T\left[k'N + n'\right] z^{-k'N} \quad (31)$$

jest określany jako (n')-ty rozkład wielofazowy $P_T(z)$. Podstawienie równania (31) do równania (30) daje

$$P_T(z) = \sum_{n'=0}^{N-1} P^{n'}_T(z^N) z^{-n'} \quad (32)$$

który jest wielofazowym rozkładem prototypowego filtra. Podobnie, próbkowanie $h_n(t)$ z przedziałem T_s otrzymuje

$$\begin{aligned} h_n[l] = h_n(lT_s) &= p_T[l]e^{2\pi jn\Delta flT_s + j\phi_n} \\ &= p_T[l]e^{2\pi jnl/N + j\phi_n}, \quad l = 0, 1, \ldots, \mathfrak{L} - 1 \end{aligned} \tag{33}$$

Jiang i Kaiser obliczają transformację Z n-tego filtra syntezy

$$\begin{aligned} H_n(z) &= \sum_{l=0}^{\mathfrak{L}-1} h_n[l]z^{-l} \\ &= \sum_{l=0}^{\mathfrak{L}-1} p_T[l]e^{2\pi jnl/N + j\phi_n}z^{-l} \\ &= \sum_{n'=0}^{N-1}\sum_{k'=0}^{K-1} p_T\left[k'N + n'\right] e^{2\pi jn(k'N+n')/N + j\phi_n}z^{-(k'N+n')} \\ &= e^{j\phi_n}\sum_{n'=0}^{N-1} e^{2\pi jnn'/N}z^{-n'}\sum_{k'=0}^{K-1} p_T[k'N + n']e^{2\pi jnk'}z^{-k'N} \\ &= e^{j\phi_n}\sum_{n'=0}^{N-1} e^{2\pi jnn'/N}P_T^{n'}(z^N)z^{-n'}. \end{aligned}$$

(34)

Bank filtrów syntezy składa się z N filtrów, a zatem jego transformację Z można wyrazić w postaci macierzowej jako

$$\begin{bmatrix} H_1(z) \\ H_2(z) \\ \vdots \\ H_N(z) \end{bmatrix} = \begin{bmatrix} e^{j\phi_1} & & & \\ & e^{j\phi_2} & & \\ & & \vdots & \\ & & & e^{j\phi_N} \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & W^{-1} & \cdots & W^{-N+1} \\ \vdots & \cdots & \ddots & \vdots \\ 1 & W^{-N+1} & \cdots & W^{(-N+1)^2} \end{bmatrix} \begin{bmatrix} P_T^0(z^N) \\ P_T^1(z^N)z^{-1} \\ \vdots \\ P_T^N(z^N)z^{-N+1} \end{bmatrix}, \tag{35}$$

stosując pierwotny pierwiastek N-tego stopnia z jedności $W = e^{-j2\pi/N}$. Lewy wektor oznacza obroty fazowe, macierz w środku oznacza odwrotną dyskretną transformatę Fouriera (DFT), a prawy wektor jest rozkładem wielofazowym prototypowego filtra.

Filtr banku wielonośnikowego

Zasadniczo, prototypowy filtr może być zaprojektowany tak, aby osiągnąć jak najmniejszy płat boczny za pomocą filtra banku. Ta forma transmisji wielonośnej jest znana jako Filtr banku wielonośnikowego (FBMC), wykazując doskonałą właściwość emisji poza pasmem (OOB). Następujące współczynniki dziedziny częstotliwości mogą być stosowane w celu utworzenia pożądanego prototypowego filtra z nakładającym się współczynnikiem K = 4

$$p = \left[1, 0.97196, 0.707, 0.235147\right] \tag{36}$$

Na podstawie tych współczynników odpowiedź częstotliwościowa prototypowego filtra jest uzyskiwana poprzez operację interpolacji, która jest wyrażona jako

$$P(f) = \sum_{k=-K+1}^{K-1} p_k \frac{\sin\left(\pi NK\left[f - \frac{k}{NK}\right]\right)}{NK\sin\left(\pi\left[f - \frac{k}{NK}\right]\right)}, \quad (37)$$

gdzie N jest całkowitą liczbą podnośnych, K jest współczynnikiem nakładania się, a p_k jest mapowane z wyżej wymienionych współczynników, tj. $p_0 = 1$, $p_{\pm1} = 0{,}97196$, $p_{\pm2} = 0{,}707$ i $p_{\pm3} = 0{,}235147$. Następnie jego odpowiedź impulsową $p_T(t)$ można uzyskać za pomocą odwrotnej transformacji Fouriera w następujący sposób:

$$p_T(t) = 1 + \sum_{k=1}^{K-1} p_k \cos\left(\frac{2\pi kt}{KT}\right). \quad (38)$$

Odpowiedź częstotliwościowa podnośnej FBMC jest bardzo zwarta. Można pominąć tętnienia podnośnej FBMC, a nawet nie ma ICI między dwoma niesąsiadującymi podnośnymi. Jednak ta cecha częstotliwości ma swoją cenę w domenie czasowej, gdzie filtr prototypowy obejmuje ponad K = 4 symbole, a nie prostokątny filtr prototypowy zajmujący tylko jeden symbol. Poprzez próbkowanie p_T (t) uzyskuje się filtr prototypowy w czasie dyskretnym o długości KN:

$$\begin{aligned} p_T[s] &= p_T(sT_s) \\ &= 1 + \sum_{k=1}^{K-1} p_k \cos\left(\frac{2\pi ksT_s}{KNT_s}\right) \\ &= 1 + \sum_{k=1}^{K-1} p_k \cos\left(\frac{2\pi ks}{KN}\right), \quad s = 0, 1, \ldots, KN-1. \end{aligned} \quad (39)$$

Następnie rozkład wielofazowy (n') można uzyskać przez

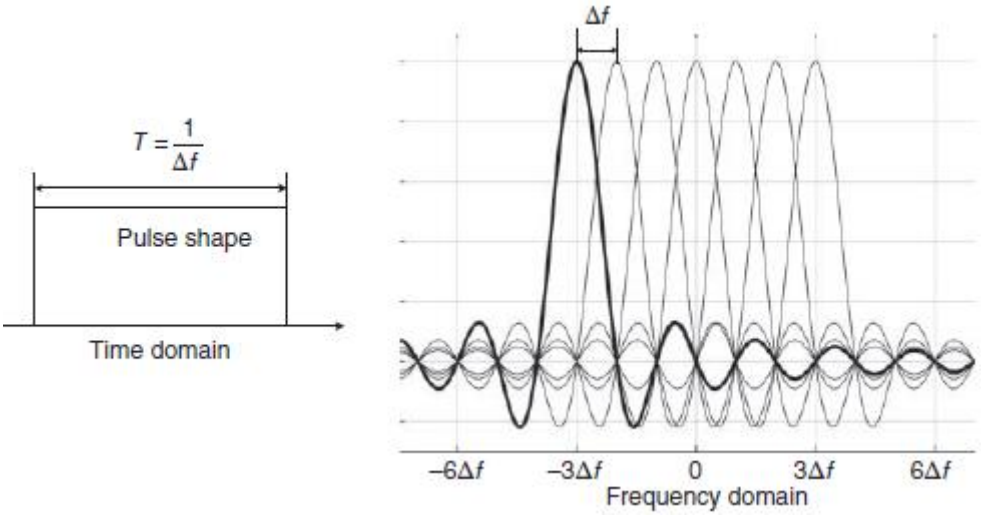
$$p_T^{n'}[k'] = p_T[k'N + n'], \quad k' = 0, 1, \ldots, K-1, \quad (40)$$

który jest w rzeczywistości filtrem Finite Impulse Response (FIR) o długości K. Ten filtr jest stosowany dla (n')-tej podnośnej FBMC, a szereg filtrów FIR N tworzy sieć wielofazową do generowania sygnałów FBMC. Dzięki wykorzystaniu DFT różnica między transmisją FBMC a transmisją z ortogonalnym zwielokrotnieniem z podziałem częstotliwości (OFDM) dotyczy jedynie implementacji wielofazowej. Stąd technologia OFDM, która jest szeroko stosowana w systemach komunikacji mobilnej i bezprzewodowej, takich jak Wireless Fidelity (Wi-Fi), 4G LTE i 5G NR, może być uważana za szczególny przypadek FBMC. W następnej sekcji omówiona zostanie zasada i kluczowe kwestie technologii OFDM jako obiecującej techniki modulacji dla nadchodzącego systemu 6G. Ortogonalne multipleksowanie z podziałem częstotliwości. Ze względu na zdolność radzenia sobie z wielodrożnym zanikaniem częstotliwości bez konieczności złożonej korekcji i prostej implementacji poprzez wykorzystanie cyfrowej transformaty Fouriera, OFDM stała się najbardziej dominującą techniką modulacji dla przewodowych i bezprzewodowych systemów komunikacyjnych w ciągu ostatnich dwóch dekad. Była szeroko stosowana w wielu znanych standardach, np. Digital Subscriber Line (DSL), Digital Video Broadcasting-Terrestrial (DVB-T), Wi-Fi, Worldwide Inter-operability for Microwave Access (WiMAX), LTE i LTE-Advanced. Po obszernym porównaniu wszystkich możliwych technik, OFDM została przyjęta jako jeden z krytycznych elementów 5G NR, jako kompleksowy kompromis między wydajnością, złożonością, porównywalnością i solidnością. Przewiduje się, że będzie ona służyć jako kluczowa

technologia w nadchodzącej transmisji 6G zarówno w konwencjonalnym paśmie sub-6 GHz, jak i pasmach wysokiej częstotliwości. Jako rodzaj modulacji wielonośnej, główne cechy transmisji OFDM, które odróżniają ją od multipleksowania z podziałem częstotliwości wielu kanałów wąskopasmowych, to:

● Zastosowanie typowo bardzo dużej liczby ortogonalnych podnośnych, a nie tylko kilku nienakładających się na siebie nośnych.

● Zastosowanie prostego prostokątnego kształtowania impulsów i widma podnośnych o kształcie sinusoidalnym, jak pokazano na rysunku.



● Ścisłe upakowanie podnośnych w dziedzinie częstotliwości z odstępem między podnośnymi wynoszącym Δf = 1⁄T, gdzie T jest czasem trwania okresu symbolu.

Schemat OFDM można uznać za szczególny przypadek FMBC, w którym prostokątny prototypowy filtr modelowany przez

$$p_0(t) = \begin{cases} 1, & -\frac{T}{2} \leqslant t < \frac{T}{2} \\ 0, & \text{others} \end{cases} \quad (41)$$

jest stosowany , jak pokazano na rysunku. Próbkując z przedziałem $T_s = T/N$, dyskretny prototyp filtru prostokątnego można podać za pomocą:

$$p_0[l] = \begin{cases} 1, & 0 \leqslant l < N-1 \\ 0, & \text{others} \end{cases} \quad (42)$$

Transformata Fouriera prostokątnego prototypowego filtra to

$$P_0(f) = \int_{-\infty}^{\infty} p_0(t) e^{-2\pi j f t} dt = \int_{-T/2}^{T/2} e^{-2\pi j f t} dt = \frac{\sin \pi f T}{\pi f} = T \text{sinc}\left(\frac{f}{\Delta f}\right) \quad (43)$$

implikując widmo sinc-kształtne, które osiąga ortogonalność podnośnej w dziedzinie częstotliwości. W związku z tym filtry syntezy w równaniu (19) i filtry analizy w równaniu (22) stają się

$$h_n(t) = \begin{cases} e^{2\pi jn\Delta ft + j\phi_n}, & -\frac{T}{2} \leqslant t < \frac{T}{2} \\ 0, & \text{others} \end{cases} \quad (44)$$

$$g_k(t) = \begin{cases} e^{-(2\pi jk\Delta ft + j\phi_k)}, & -\frac{T}{2} \leqslant t < \frac{T}{2} \\ 0, & \text{others} \end{cases} \quad (45)$$

i

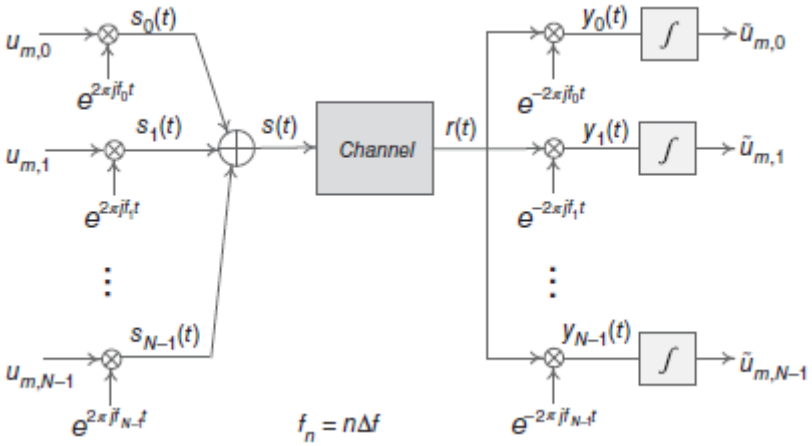
odpowiednio. Ponieważ obrót fazy nie wpływa na ortogonalność podnośnych OFDM, możemy pominąć fazę i użyć $\phi_n = 0$, $\forall n = 1, 2, \ldots, N$. W złożonej notacji pasma podstawowego podstawowy sygnał OFDM s(t) podczas okresu symbolu $mT \leqslant t < (m + 1)T$ można zatem podać za pomocą

$$s(t) = \sum_{n=0}^{N-1} s_n(t) = \sum_{n=0}^{N-1} u_{m,n} e^{2\pi jn\Delta ft}, \quad (46)$$

gdzie $s_n(t)$ jest n-tą zmodulowaną podnośną o częstotliwości $f_n = n \Delta f$. Transformatę Fouriera n-tej niemodulowanej podnośnej można obliczyć za pomocą

$$\begin{aligned} P_n(f) &= \int_{-\infty}^{\infty} p_0(t) e^{2\pi jn\Delta ft} e^{-2\pi jft} dt \\ &= \int_{-T/2}^{T/2} e^{2\pi jn\Delta ft} e^{-2\pi jft} dt \\ &= T\text{sinc}\left(\frac{f}{\Delta f} - n\right). \end{aligned} \quad (47)$$

Oznacza to, że widmo n-tej podnośnej można uzyskać po prostu przesuwając P0(f) na osi częstotliwości o przesunięcie nΔf , tj. $P_n(f) = P_0(f - n\Delta f)$, jak pokazano na rysunku powyżej. Podstawowe zasady modulacji i demodulacji OFDM można przedstawić na rysunku.



Demodulacja OFDM wykorzystuje bank korelatorów, po jednym dla każdej podnośnej. Przy ortogonalności między podnośnymi, dwie dowolne podnośne OFDM nie powodują żadnych zakłóceń w idealnym przypadku, nawet jeśli widmo sąsiednich podnośnych się nakłada. Tak więc unikanie zakłóceń między podnośnymi OFDM nie wynika po prostu z separacji częstotliwości, co ma miejsce w przypadku multipleksowania z podziałem częstotliwości. Zamiast tego ortogonalność podnośnej wynika ze specyficznej struktury domeny częstotliwości każdej podnośnej w połączeniu ze szczególnym wyborem odstępu podnośnej równego szybkości symboli, tj. $\Delta f = 1/T$. Pomijając upośledzenia kanału

i szum addytywny, tj. r(t) = s(t), i przechodząc przez odpowiedni filtr analizy $g_k(t)$, wynikowy sygnał na k-tej podnośnej podczas m-tego okresu symboli to

$$y_k(t) = \sum_{n=0}^{N-1} u_{m,n} e^{2\pi jn\Delta ft} e^{-2\pi jk\Delta ft}$$
$$= \sum_{n=0}^{N-1} u_{m,n} e^{2\pi j(n-k)\Delta ft}.$$

Całkowanie w okresie jednego symbolu daje

$$\tilde{u}_{m,k} = \frac{1}{T}\int_{mT}^{(m+1)T} y_k(t)dt$$
$$= \frac{1}{T}\int_{mT}^{(m+1)T} \sum_{n=0}^{N-1} u_{m,n} e^{2\pi j(n-k)\Delta ft} dt$$
$$= \frac{1}{T}\int_{mT}^{(m+1)T} u_{m,n}dt + \underbrace{\frac{1}{T}\sum_{n=0,n\neq k}^{N-1} u_{m,n}\int_{mT}^{(m+1)T} e^{2\pi j(n-k)\Delta ft}dt}_{\text{Inter-carrier interference}}.$$

Ze względu na ortogonalność między podnośnymi, ICI można w zasadzie wyeliminować, a my mamy

$$\tilde{u}_{m,k} = u_{m,n}, \quad \forall n = k. \tag{48}$$

W rezultacie symbol niosący informację $u_{m,n}$ może zostać dostarczony przez n-tą podnośną w trakcie m-tego okresu symbolu, a łącznie N symboli niosących informację jest przesyłanych równolegle w trakcie jednego okresu symbolu.

Implementacja DFT

Chociaż bank par modulator-korelator pokazany na rysunku powyżej może być użyty do opisu podstawowych zasad modulacji i demodulacji OFDM, nie jest to najbardziej odpowiednia struktura do rzeczywistej implementacji. W rzeczywistości, ze względu na swoją specyficzną konstrukcję i szczególny wybór odstępu między podnośnymi, OFDM umożliwia implementację o niskiej złożoności wykorzystującą przetwarzanie DFT. Może być dalej implementowany przez wydajny obliczeniowo algorytm o nazwie Szybka Transformata Fouriera (FFT), jeśli liczba podnośnych jest potęgą 2. Ze względu na prostokątny kształt impulsu symbole OFDM nie nakładają się w dziedzinie czasu (tj. współczynnik nakładania się K = 1). W rezultacie rozkład wielofazowy zdefiniowany w równaniu (31) jest uproszczony do

$$P_0^{n'}(z^N) = \sum_{k'=0}^{K-1} p_0[k'N + n']z^{-k'N}$$
$$= \sum_{k'=0}^{0} p_0[n']z^0$$
$$= 1, \quad n' = 0, 1, \ldots, N-1. \tag{49}$$

W ten sposób równanie (35) można zapisać w postaci

$$\begin{bmatrix} H_1(z) \\ H_2(z) \\ \vdots \\ H_N(z) \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & W^{-1} & \cdots & W^{-N+1} \\ \vdots & \cdots & \ddots & \vdots \\ 1 & W^{-N+1} & \cdots & W^{(-N+1)^2} \end{bmatrix}, \quad (50\}$$

co jest równoważne z odwrotną dyskretną transformatą Fouriera (IDFT), co oznacza, że sygnał OFDM może być generowany przez modulator IDFT. Ponieważ Bw = N Δ f można postrzegać jako nominalną szerokość pasma transmisji OFDM, dyskretny sygnał OFDM można uzyskać przez próbkowanie s(t) w równaniu (46) z szybkością fs ⩾ N Δ f, aby spełnić twierdzenie o próbkowaniu. Zakładając, że szybkość próbkowania jest wielokrotnością odstępu między podnośnymi, tj. $f_s = N_s \Delta f$, wiemy, że $N_s \geqslant N$, co w OFDM nazywa się nadpróbkowaniem. Na przykład transmisja LTE obsługuje około N = 1200 podnośnych, podczas gdy rozmiar DFT jest wybrany jako N_s = 2048. Odpowiada to częstotliwości próbkowania $f_s = N_s \Delta f$ = 30,72 MHz, podanej Δ f = 15 kHz w LTE. Pozostałe 848 podnośnych nie przenosi niczego, przypisując symbol zerowy '0' i są ogólnie nazywane wirtualnymi podnośnymi OFDM. Przy tych założeniach próbkowany sygnał OFDM podczas m-tego symbolu, innymi słowy, sekwencja OFDM w czasie dyskretnym może być wyrażona jako

$$\begin{aligned} s^m[k] = s(kT_s) &= \sum_{n=0}^{N-1} s_n(kT_s) \\ &= \sum_{n=0}^{N-1} u_{m,n} e^{2\pi jn\Delta fkT_s} \\ &= \sum_{n=0}^{N-1} u_{m,n} e^{2\pi jnk/N_s}, \quad k = 0, 1, \ldots, N_s - 1. \end{aligned} \quad (51)$$

Dodając na końcu kilka wirtualnych podnośnych, możemy utworzyć alternatywne wyrażenie przesyłanych symboli

$$a_{m,n} = \begin{cases} u_{m,n}, & 0 \leqslant n < N \\ 0, & N \leqslant n < N_s \end{cases} \quad (52)$$

Następnie równanie (51) można zapisać w postaci

$$\begin{aligned} s^m[k] &= \sum_{n=0}^{N-1} u_{m,n} e^{2\pi jnk/N_s} \\ &= \sum_{n=0}^{N-1} u_{m,n} e^{2\pi jnk/N_s} + \sum_{n=N}^{N_s-1} 0 \cdot e^{2\pi jnk/N_s} \\ &= \sum_{n=0}^{N_s-1} a_{m,n} e^{2\pi jnk/N_s}, \end{aligned} \quad (53)$$

co jest równoważne z N_s-punktowym IDFT. Ujawnia, że sekwencja OFDM w czasie dyskretnym jest dokładnie równa DFT symboli modulacji, po której następuje konwersja cyfrowo-analogowa. OFDM jest rodzajem transmisji blokowej. Blok symboli modulacji $u_m = [u_{m,0}, u_{m,1}, \ldots, u_{m,N-1}]^T$ jest najpierw rozszerzany o zera do długości Ns, co daje

$$\mathbf{a}_m = [u_{m,0}, u_{m,1}, \ldots, u_{m,N-1}, 0, \ldots, 0]^T. \quad (54)$$

Wykonaj przetwarzanie IDFT na AM, aby uzyskać sekwencję OFDM

$$\mathbf{s}_m = [s_m[0], s_m[1], \ldots, s_m[N_s - 1]]^T. \quad (55)$$

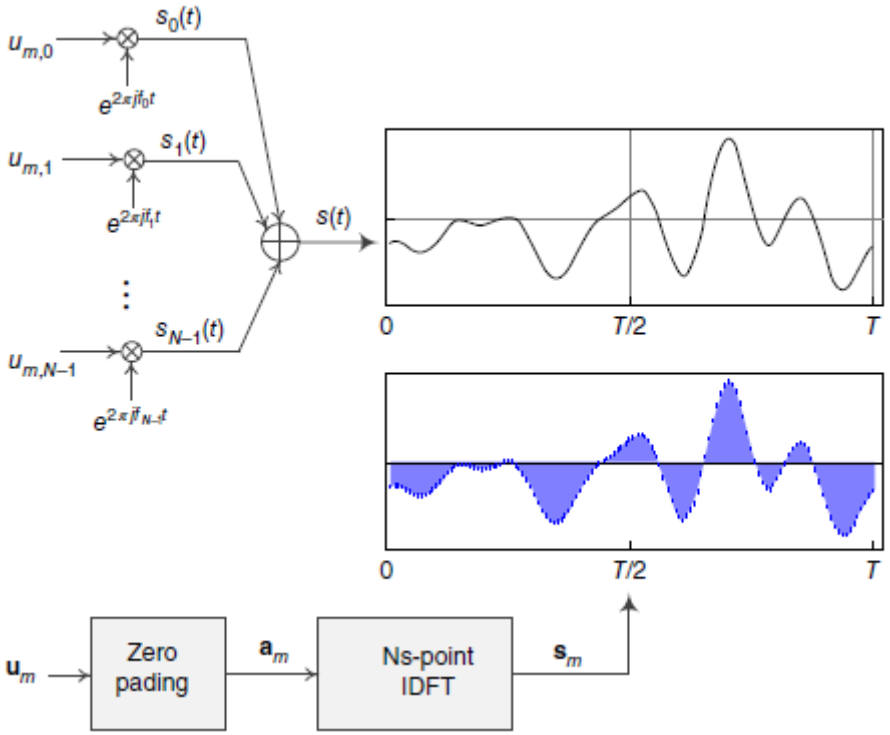
Definiujemy macierz

$$\mathbf{F} = \begin{pmatrix} \omega_{N_s}^{0\cdot 0} & \omega_{N_s}^{0\cdot 1} & \cdots & \omega_{N_s}^{0\cdot(N_s-1)} \\ \omega_{N_s}^{1\cdot 0} & \omega_{N_s}^{1\cdot 1} & \cdots & \omega_{N_s}^{1\cdot(N_s-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{N_s}^{(N_s-1)\cdot 0} & \omega_{N_s}^{(N_s-1)\cdot 1} & \cdots & \omega_{N_s}^{(N_s-1)\cdot(N_s-1)} \end{pmatrix} \quad (56)$$

przy pierwotnym pierwiastku N-tego s z jedności $\omega_{Ns} = e^{-2\pi j/Ns}$ modulację OFDM (czyli IDFT) można zapisać w postaci macierzowej jako

$$\mathbf{s}_m = \mathbf{F}^* \mathbf{a}_m = N_s \mathbf{F}^{-1} \mathbf{a}_m, \quad (57)$$

gdzie indeks górny $(\cdot)^*$ oznacza sprzężenie zespolone, a indeks górny $(\cdot)^{-1}$ oznacza odwrotność macierzy kwadratowej. Jak pokazano na rysunku ,



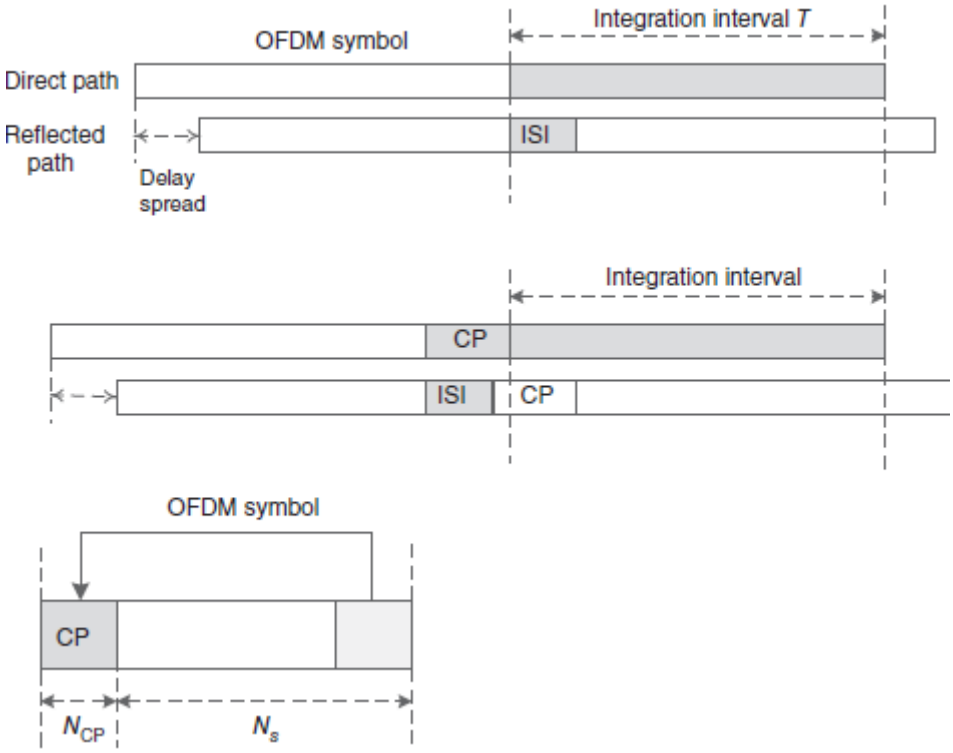
sygnał OFDM, składający się z N modulowanych podnośnych, ma taką samą obwiednię jak sekwencja OFDM uzyskana poprzez N_s-punktową DFT symboli modulacji. Podobnie przetwarzanie DFT można wykorzystać do demodulacji OFDM, zastępując bank N dekorelatorów próbkowaniem z szybkością $f_s = N_s \Delta f$, po którym następuje operacja Ns-punktowa DFT/FFT.

Prefiks cykliczny

Poprzednia podsekcja zilustrowała, że nieuszkodzony sygnał OFDM można zdemodulować w odbiorniku bez zakłóceń między podnośnymi Ortogonalność jest osiągana, ponieważ każda podnośna ma całkowitą liczbę okresów wykładników zespolonych w ciągu jednego okresu symbolu OFDM, matematycznie oznaczoną jako

$$e^{2\pi j f_n t} = e^{2\pi j n \Delta f t} = e^{2\pi j \frac{nt}{T}}, -\frac{T}{2} \leqslant t < \frac{T}{2}. \quad (58)$$

W praktyce jednak ortogonalność ta może zostać utracona w kanale dyspersyjnym w czasie. Dzieje się tak, ponieważ przedział korelacji dla jednej ścieżki będzie nachodził na granicę symboli innej ścieżki. Przedział całkowania niekoniecznie będzie odpowiadał całkowitej liczbie okresów wykładników zespolonych, ponieważ symbole modulacji mogą się różnić między dwoma kolejnymi symbolami. W rezultacie kanał dyspersyjny w czasie powoduje nie tylko interferencję między symbolami, ale także interferencję między nośnymi w transmisji OFDM, jak pokazano na rysunku.



Aby rozwiązać ten problem i uczynić sygnał OFDM odpornym na dyspersję czasową, Peled i Ruiz [1980] zaproponowali prefiks cykliczny (CP) lub pierwotnie nazywany rozszerzeniem cyklicznym, odnoszącym się do prefiksu symbolu OFDM. Wstawianie prefiksu cyklicznego oznacza, że ostatnia część symbolu OFDM jest kopiowana i wstawiana na początku symbolu OFDM. Wstawianie zwiększa zatem długość symbolu OFDM z T do To = T + T_{CP}, gdzie T_{CP} jest długością prefiksu cyklicznego, co prowadzi do zmniejszenia szybkości symbolu OFDM. Jeśli interwał korelacji w demodulatorze jest nadal realizowany w okresie symbolu T, ortogonalność podnośnej może być zachowana tak długo, jak długo czas trwania CP jest większy niż rozproszenie opóźnienia. W praktyce wstawianie CP jest realizowane w dyskretnej czasowo sekwencji OFDM, gdzie ostatnie próbki NCP są kopiowane i wstawiane na początku bloku, zwiększając długość bloku z N_s do N_s + N_{CP}. W demodulatorze odpowiednie próbki są odrzucane przed demodulacją DFT. Przyjmuje się, że dyskretna odpowiedź impulsowa kanału zanikającego selektywnie pod względem częstotliwości w czasie m wynosi

$$\mathbf{h}_m = \left[h_m[0], h_m[1], \ldots, h_m[\mathfrak{L}-1]\right]^T, \quad (59)$$

gdzie $\mathfrak{L}$ oznacza długość filtru kanałowego. Gdy m-ty symbol OFDM sm przechodzi przez kanał, wynikowy sygnał w notacji pasma podstawowego w czasie dyskretnym jest obliczany przez Oppenheima i innych.

$$\mathbf{r}_m = \mathbf{h}_m * \mathbf{s}_m. \quad (60)$$

gdzie $*$ oznacza splot liniowy. Długość sygnału wyjściowego wynosi $\mathfrak{L} + N_s - 1$, co jest dłuższe od sygnału wejściowego, ponieważ długość filtru kanałowego jest większa niż jeden w kanale zanikającym selektywnie pod względem częstotliwości. W związku z tym odebrany sygnał można oznaczyć jako

$$\mathbf{r}_m = \left[r_m[0], r_m[1], \cdots, r_m[Q-1]\right]^T \quad (61)$$

gdzie $Q = \mathfrak{L} + N_s - 1$, i

$$r_m[q] = \sum_{l=0}^{\mathfrak{L}-1} h_m[l] s_m[q-l], \quad q = 0, 1, \ldots, Q-1 \quad (62)$$

Bez CP próbki resztkowe o długości $\mathfrak{L} - 1$ na końcu każdego symbolu OFDM nakładają się na jego kolejny symbol OFDM, powodując ISI i niszcząc ortogonalność podnośnej. Intuicyjnie, wstawienie odstępu ochronnego między dwoma kolejnymi symbolami OFDM może pochłonąć resztkę poprzedniego symbolu OFDM. Biorąc pod uwagę równanie (55), sekwencję OFDM z wstawieniem CP można wyrazić wzorem

$$\mathbf{x}_m = \underbrace{\left[s_m[N_s - N_{\mathrm{CP}}], \ldots, s_m[N_s-1]\right]}_{\text{CP Insertion}}, s_m[0], s_m[1], \ldots, s_m[N_s-1]\Big]^T. \quad (63)$$

Splot x_m z h_m daje

$$\mathbf{y}_m = \mathbf{h}_m * \mathbf{x}_m. \quad (64)$$

Długość y_m wynosi $S = \mathfrak{L} + N_s + N_{CP} - 1$, a następnie możemy zapisać otrzymany sygnał w postaci

$$\mathbf{y}_m = \left[y_m[0], y_m[1], \ldots, y_m[S-1]\right] \quad (65)$$

którego wpis jest równy

$$y_m[s] = \sum_{l=0}^{\mathfrak{L}-1} h_m[l] x_m[s-l]. \quad (66)$$

Dopóki rozpiętość rozproszenia opóźnienia nie przekracza długości CP, interferencja między symbolami może zostać pochłonięta, podczas gdy interferencja między nośnymi może zostać również uniknięta, ponieważ ortogonalność podnośnych jest zachowana podczas interwału integracji. Wadą wstawiania cyklicznego prefiksu jest utrata mocy i szerokości pasma, gdy szybkość symboli OFDM maleje. Jednym ze sposobów zminimalizowania takiej straty jest zmniejszenie odstępu między podnośnymi Δf, co w konsekwencji powoduje odpowiedni wzrost okresu symbolu T. Jednak zwiększy

to wrażliwość transmisji OFDM na szybkie wahania kanału z powodu dużego rozproszenia Dopplera. Istotne jest również zrozumienie, że CP niekoniecznie musi obejmować całą długość dyspersji czasowej kanału. Zasadniczo istnieje kompromis między utratą mocy a uszkodzeniem sygnału (interferencja między symbolami i między podnośnymi) z powodu resztkowej dyspersji czasowej nieobjętej przez cykliczny prefiks.

Przetwarzanie sygnału w dziedzinie częstotliwości

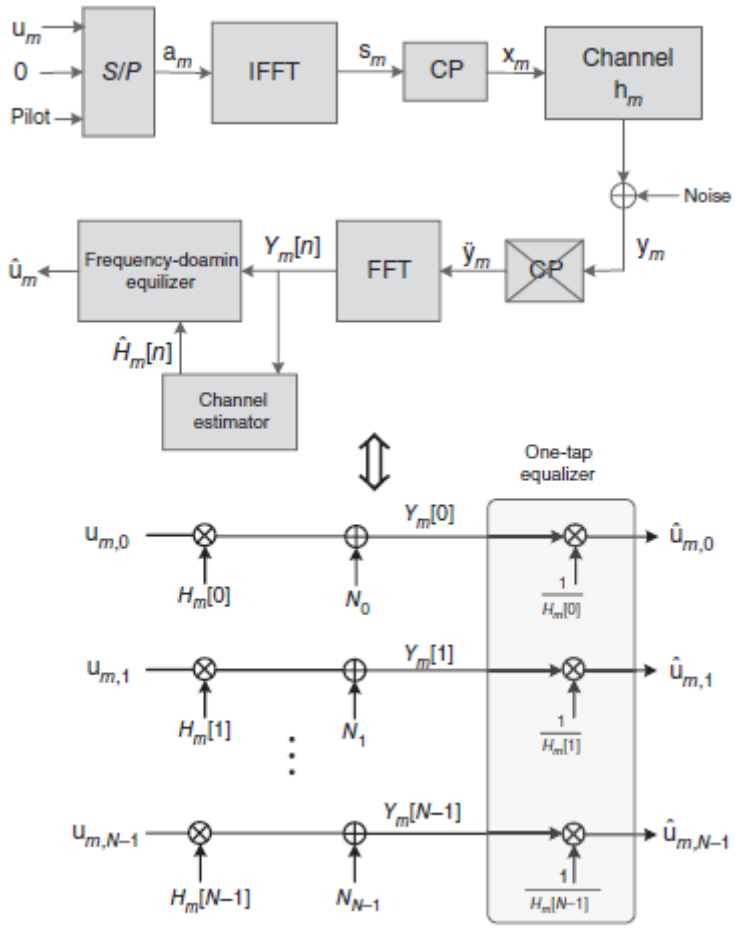
Oprócz

● wyeliminowania ISI z poprzedniego symbolu i

● zachowania ortogonalności podnośnej, zastosowanie prefiksu cyklicznego może

● przekształcić splot liniowy z filtrem kanałowym w splot kołowy, znany również jako splot cykliczny, umożliwiając proste przetwarzanie sygnału w dziedzinie częstotliwości.

Zakładając wystarczająco duży prefiks cykliczny, splot liniowy kanału radiowego o dyspersji czasowej pojawi się jako splot kołowy podczas interwału integracji demodulatora. Połączenie modulacji OFDM, kanału radiowego o dyspersji czasowej i demodulacji OFDM można następnie postrzegać jako zestaw równoległych podkanałów w dziedzinie częstotliwości, jak pokazano na rysunku.



W odbiorniku CP jest odrzucane, a do demodulacji używane są tylko próbki w przedziale całkowania od N_{CP} do $N_{CP} + N_s - 1$. Matematycznie rzecz biorąc, wejście demodulatora DFT to

$$\ddot{\mathbf{y}}_m = \left[\ddot{y}_m[0], \ddot{y}_m[1], \ldots, \ddot{y}_m[N_s - 1]\right]^T = \left[y_m[N_{\mathrm{CP}}], y_m[N_{\mathrm{CP}} + 1], \ldots, y_m[N_{\mathrm{CP}} + N_s - 1]\right]^T, \quad (67)$$

wyodrębniono z y_m w równaniu (65). Niesamowite jest to, że wyodrębniona sekwencja jest dokładnie równoważna splotowi kołowemu s_m i rozszerzonemu wypełnieniu filtru kanałowego z zerami do długości N_s, tj.

$$\ddot{\mathbf{y}}_m = \mathbf{s}_m \otimes \mathbf{h}_{N_s}, \quad (68)$$

i alternatywnie

$$\ddot{y}_m[k] = y_m[N_{\mathrm{CP}} + k] = \sum_{i=0}^{N_s-1} s_m[i] h_{N_s}\left[(k - i)_{N_s}\right], \quad \forall k = 0, \ldots, N_s - 1, \quad (69)$$

gdzie $\otimes$ oznacza splot kołowy, $(\cdot)_{N_s}$ oznacza okresowe przesunięcie z N_s,

i

$$\mathbf{h}_{N_s} = \left[h_{N_s}[0], h_{N_s}[1], \ldots, h_{N_s}[N_s - 1]\right]^T = \left[h_m[0], h_m[1], \ldots, h_m[\mathfrak{L} - 1], \underbrace{0, \ldots, 0}_{N_s - \mathfrak{L}}\right]^T. \quad (70)$$

Szczególną zaletą formowania splotu kołowego w transmisji OFDM jest umożliwienie prostego przetwarzania w dziedzinie częstotliwości i korekcji jednym dotknięciem. Zgodnie z teorią przetwarzania sygnałów, splot kołowy dwóch skończonych sekwencji o tej samej długości w dziedzinie czasu, a nie splot liniowy, odpowiada mnożeniu ich DFT w dziedzinie częstotliwości. Matematycznie, jeśli $\ddot{y}_m = s_m \otimes h_{Ns}$, mamy

$$Y_m[n] = H_m[n] X_m[n], \quad \forall n = 0, 1, \ldots, N_s - 1, \quad (71)$$

gdzie $Y_m[n]$ i $X_m[n]$ to n-ty wpis sekwencji odebranych i przesłanych w dziedzinie częstotliwości, które są odpowiednio DFT $\ddot{y}_m$ i s_m, co daje

$$Y_m[n] = \sum_{k=0}^{N_s-1} \ddot{y}_m[k] e^{-2\pi jnk/N_s} = \sum_{k=0}^{N_s-1} y_m[k + N_{\mathrm{CP}}] e^{-2\pi jnk/N_s}, \quad (72)$$

i

$$X_m[n] = \sum_{k=0}^{N_s-1} s_m[k] e^{-2\pi jnk/N_s}. \quad (73)$$

Przypominając równanie (57), ciąg OFDM sm jest IDFT symboli niosących informacje

$$\mathbf{a}_m = \left[u_{m,0}, u_{m,1}, \ldots, u_{m,N-1}, 0, \ldots, 0\right]^T \quad (74)$$

i dlatego mamy

$$X_m[n] = a_m[n], \quad \forall n = 0, 1, \ldots, N_s - 1, \quad (75)$$

i

$$X_m[n] = u_{m,n}, \quad \forall n = 0, 1, \ldots, N - 1. \quad (76)$$

Biorąc pod uwagę szum w dziedzinie częstotliwości oznaczony jako $N_m[n]$, równanie (71) można zapisać jako

$$Y_m[n] = H_m[n]u_{m,n} + N_m[n], \quad \forall n = 0, 1, \ldots, N - 1, \quad (77)$$

co oznacza, że transmisję OFDM można traktować jako zbiór równoległych podkanałów w dziedzinie częstotliwości, jak pokazano na rysunku . Złożone odczepy kanału w dziedzinie częstotliwości $H_m[n]$, n = 0, 1,…,N_s − 1 można uzyskać, wykonując DFT na h_{Ns}, tj.

$$H_m[n] = \sum_{l=0}^{N_s-1} h_{N_s}[l]e^{-2\pi jnl/N_s} = \sum_{l=0}^{g-1} h_m[l]e^{-2\pi jnl/N_s} \quad (78)$$

Ponieważ modulacja i demodulacja są przeprowadzane niezależnie na każdej podnośnej, szacowanie i wyrównywanie kanału staje się prostsze w dziedzinie częstotliwości. Załóżmy, że P[n] jest znanym symbolem transmisji w odbiorniku, nazywanym pilotem, możemy oszacować odpowiedź kanału w punkcie wstawienia pilota jako

$$\hat{H}_p[n] = \frac{Y_p[n]}{P[n]}, \quad (79)$$

gdzie $Y_p[n]$ jest odebranym sygnałem odpowiadającym P[n]. Na podstawie oszacowania kanału przy pilocie $\hat{H}_p[n]$, odpowiedzi kanału przy wszystkich podnośnych $\hat{H}_p[n]$, można oszacować poprzez operację interpolacji. W ten sposób odzyskiwanie transmitowanego symbolu można zrealizować za pomocą prostego korektora jednopunktowego

$$\hat{X}[n] = \frac{Y[n]}{\hat{H}[n]}. \quad (80)$$

Do szacowania kanału można stosować zaawansowane algorytmy, od prostego uśredniania w połączeniu z interpolacją liniową po szacowanie MMSE, polegające na bardziej szczegółowej wiedzy o charakterystyce kanału w dziedzinie czasu i częstotliwości.

Emisja poza pasmem

Ze względu na duże płaty boczne, które zanikają asymptotycznie z f−2, upływ mocy OOB sygnałów OFDM jest niedopuszczalnie wysoki dla praktycznych systemów. Zakłócenia sygnałów OFDM na sąsiednich kanałach wynoszą około −20 dB, znacznie więcej niż wymaganie współczynnika mocy zakłóceń sąsiedniego kanału (ACIR) wynoszącego −45 dB określonego w 3GPP LTE. Ponadto w

przypadku niektórych konkretnych scenariuszy wdrożeniowych, takich jak sieci oparte na radiu kognitywnym w pasmach telewizji białej, ACIR zdefiniowany przez Federalną Komisję Łączności (FCC) jest znacznie obniżony do −72 dB. Ze względu na niską złożoność implementacji, wstawianie pasm ochronnych jest często wykorzystywane w systemach OFDM w celu zminimalizowania wycieku mocy. Chociaż cała szerokość pasma kanału jest przydzielona do dedykowanego kanału LTE, widmo to nie może być w pełni wykorzystane do przesyłania sygnałów. Pasma ochronne są wstawiane poprzez dezaktywację podnośnych leżących na krawędziach pasma widmowego, stosując wirtualne podnośne OFDM. Wykorzystanie pasm ochronnych w jakiś sposób zmniejsza ilość wycieku mocy OOB, ale nieuchronnie wiąże się z utratą wydajności widmowej. Szerokość pasma kanału to ilość zasobów widmowych przydzielonych do dedykowanego systemu, podczas gdy szerokość pasma transmisji to szerokość widma, która jest faktycznie zajmowana przez sygnały transmisji. Oczywiście szerokość pasma transmisji nie może być większa niż szerokość pasma kanału. W LTE termin Resource Block (RB) jest definiowany jako zestaw podnośnych OFDM, co odpowiada 12 podnośnym rozciągającym się na szerokość pasma sygnału 180 kHz. Szerokość pasma transmisji w LTE jest parametryzowana przez liczbę RB. Na przykład kanał 1,4 MHz jest w stanie przesłać do 6 RB, co odpowiada szerokości pasma sygnału 1,08 MHz. Różnica między szerokością pasma kanału a szerokością pasma transmisji jest dokładnie taka sama jak szerokość pasm ochronnych. Aby dać ilościową ocenę utraty wydajności widmowej, parametry związane z pasmami ochronnymi określonymi w 3GPP LTE zostały podsumowane w Tabeli 1 jako przykład.

Channel bandwidth (MHz)	1.4	3	5	10	15	20
Number of RBs	6	15	25	50	75	100
Transmission bandwidth (MHz)	1.08	2.7	4.5	9	13.5	18
Guard band (MHz)	0.32	0.3	0.5	1	1.5	2
Spectral loss	22.85%	10%	10%	10%	10%	10%

Z tej tabeli możemy wywnioskować, że utrata wydajności widmowej spowodowana wykorzystaniem pasm ochronnych wynosi ponad 10% w systemie LTE. Oprócz wstawiania pasm ochronnych, zaawansowane algorytmy przetwarzania sygnału, takie jak okienkowanie w dziedzinie czasu, aktywna eliminacja zakłóceń, ważenie podnośnych, wstępne kodowanie widmowe i filtrowanie dolnoprzepustowe, zostały zaprojektowane w celu tłumienia wycieku mocy sygnałów OFDM. Okienkowanie w dziedzinie czasu Ten schemat stosuje odpowiednie okna, takie jak półsinus lub okno Hanninga, do przesyłanego sygnału. Poprzez wygładzanie amplitudy sygnału do zera na granicach symboli, płaty boczne mogą być znacząco ograniczone. Różne funkcje okienkowania można jednolicie formułować za pomocą

$$w(t) = R(t/T) * g(t), \quad (81)$$

gdzie R(t) oznacza znormalizowany impuls prostokątny

$$R(t) = \begin{cases} 1, & 0 \leqslant t < 1 \\ 0, & \text{otherwise} \end{cases} \quad (82)$$

Najczęściej stosowany impuls półsinusoidalny jest definiowany jako

$$g(t) = \frac{\pi}{2\beta T} \sin\left(\frac{\pi t}{\beta T}\right) R(t/\beta T) \quad (83)$$

gdzie β reprezentuje współczynnik odchylenia. Należy zauważyć, że $\beta > 0$ skutecznie wydłuża czas trwania symbolu OFDM, zwiększając w ten sposób narzut i obniżając wydajność widmową. Nawet jeśli użyto małej wartości β, schemat okienkowania może osiągnąć znaczne tłumienie płatów bocznych. Co więcej, złożoność obliczeniowa spowodowana okienkowaniem każdego symbolu OFDM jest pomijalna w porównaniu z przetwarzaniem DFT. Aktywna redukcja interferencji. Innym rozwiązaniem w celu obniżenia płatów bocznych sygnałów OFDM jest wstawienie dodatkowych podnośnych redukcji na krawędziach widma OFDM. Symbole informacyjne przesyłane podczas n-tego symbolu to $u_m = [u_{m,0}, u_{m,1},..., u_{m,N-1}]^T$. Wstawienie Nc dodatkowych symboli złożonych $g_m = [g_{m,1}, g_{m,2},..., g_{m,Nc}]^T$ skutkuje

$$\mathbf{a}_m = \left[g_{m,1}, \cdots, g_{m,\frac{N_c}{2}}, u_{m,0}, \cdots, u_{m,N-1}, g_{m,\frac{N_c}{2}+1}, \cdots, g_{m,N_c}\right]^T. \quad (84)$$

Aby zmierzyć upływ mocy, L wybranych punktów obserwacji częstotliwości jest wybieranych w zakresie widma OOB. Definiując $f_{l,nc}$ jako wkład n-tej podnośnej kasującej c do l-tego punktu obserwacji, można utworzyć macierz $L \times N_c$ C z wpisami $f_{l,nc}$. Oznaczmy przez fm wektor L-wymiarowy zawierający wkłady podnośnych niosących u_m do L punktów obserwacji. Następnie problem optymalizacji można sformułować jako liniowy problem najmniejszych kwadratów, tj.

$$\mathbf{g}_m = \arg\min_{\tilde{\mathbf{g}}_m} \left\|\mathbf{f}_m + \mathbf{C}\tilde{\mathbf{g}}_m\right\| \quad (85)$$

Rozwiązanie tego problemu sprowadza się do znalezienia optymalnego wektora ważenia gm, który minimalizuje wyciek mocy OOB. Aby uprościć proces optymalizacji, symbole $g_{m,nc}$ można ograniczyć do wstępnie zdefiniowanego zestawu symboli ilościowych. Proces optymalizacji w równaniu (85) można następnie zrealizować, stosując wyczerpujące wyszukiwanie.

Ważenie podnośnej. Ważenie podnośnej mnoży każdy symbol informacyjny przez zoptymalizowaną wagę jako

$$\hat{u}_{m,n} = g_{m,n}u_{m,n}, \quad n = 0, \ldots, N-1. \quad (86)$$

Wektor ważenia można uzyskać rozwiązując równanie optymalizacji

$$\mathbf{g}_m = \arg\min_{\tilde{\mathbf{g}}_m} \|\mathbf{S}\tilde{\mathbf{g}}_m\|^2, \quad (87)$$

gdzie S jest macierzą L × N, której elementy Sl,n odzwierciedlają składową widmową n-tej podnośnej do l-tego punktu obserwacji częstotliwości, jak zdefiniowano w aktywnej eliminacji interferencji. Zarówno wagi dla podnośnych eliminacji w aktywnej eliminacji interferencji, jak i dla symboli informacyjnych w tym schemacie są zależne od danych i dlatego muszą być określone przez rozwiązanie problemów optymalizacji ograniczonej dla każdego symbolu OFDM. Ceną za obie metody jest zwiększony stosunek mocy szczytowej do średniej, a także zmniejszony średni stosunek sygnału do szumu plus interferencja, co skutkuje pogorszeniem wydajności współczynnika błędów bitowych (BER).

Wstępne kodowanie widmowe. Aby uniknąć iteracyjnego obliczania wag na podstawie symbolu OFDM, w Jiang i Zhao zaproponowano podejście niezależne od danych zwane wstępnym kodowaniem widmowym, w którym przesyłane symbole są wstępnie kodowane przed modulacją OFDM jako

$$\tilde{\mathbf{a}}_k = \mathbf{G}\mathbf{a}_k \quad (88)$$

Matryca prekodowania G ma na celu uczynienie sygnału OFDM i jego pierwszych n pochodnych ciągłymi w fazie i amplitudzie. Oznaczając przez $x_m(t)$, $mT \leqslant t < (m + 1)T$ m-ty symbol OFDM z CP, płaty boczne można drastycznie zmniejszyć, czyniąc dwa kolejne symbole OFDM ciągłymi w ich pierwszych n pochodnych, co wyraża się przez Chunga

$$\left.\frac{d^n}{dt^n}x_{m-1}(t)\right|_{t=mT} = \left.\frac{d^n}{dt^n}x_m(t)\right|_{t=mT}. \quad (89)$$

Zwiększając rząd pochodnych n, BER ulega degradacji, ponieważ kodowanie wstępne spowoduje nierównomierny rozkład sygnału na podnośne. Aby to zrekompensować, można zastosować dekodowanie iteracyjne, co wiązałoby się z dodatkową złożonością obliczeniową. Wybór odpowiedniego n będzie kwestią pożądanego kompromisu między redukcją upływu mocy OOB a złożonością. Filtrowanie dolnoprzepustowe Inną skuteczną metodą w praktyce jest filtrowanie dolnoprzepustowe, w którym przesyłany sygnał jest filtrowany przed konwersją cyfrowo-analogową

$$\bar{s}[n] = s[n] * f[n] \quad (90)$$

gdzie s[n] oznacza sekwencję OFDM, a f [n] oznacza filtr FIR. Płat boczny sygnałów OFDM można znacznie stłumić, projektując filtr z dużym tłumieniem OOB. Na przykład filtr dolnoprzepustowy z 88 odczepami może osiągnąć tłumienie 50 dB [Jiang i Schellmann.

Ortogonalny dostęp z podziałem częstotliwości

Sieć komórkowa musi obsługiwać wielu aktywnych abonentów jednocześnie w skończonej ilości zasobów czasowo-częstotliwościowych. Dlatego też efektywny przydział zasobów radiowych pomiędzy użytkownikami jest krytycznym aspektem projektowym zarówno kanałów łącza w górę (UL), jak i łącza w dół (DL), ponieważ przepustowość jest zwykle ograniczona i droga. Udział kanału komunikacyjnego pomiędzy wieloma użytkownikami, którzy są rozproszeni geograficznie, jest określany jako dostęp wielokrotny. Najczęstsze techniki dostępu wielokrotnego, które ortogonalnie lub nieortogonalnie dzielą wymiary sygnalizacji na kanały, a następnie przypisują te kanały różnym użytkownikom, obejmują dostęp wielokrotny z podziałem czasu (TDMA), dostęp wielokrotny z podziałem częstotliwości (FDMA), dostęp wielokrotny z podziałem kodu (CDMA) i dostęp wielokrotny z podziałem przestrzeni (SDMA). Jako rozszerzenie transmisji OFDM do wdrożenia systemu wielodostępnego, dostęp wielokrotny z podziałem częstotliwości (OFDMA) zapewnia wydajną i elastyczną technologię dostępu wielokrotnego w siatce zasobów czasowo-częstotliwościowych.

Ortogonalny dostęp wielokrotny z podziałem częstotliwości

Dyskusja w poprzedniej sekcji zakładała domyślnie, że transmisja OFDM jest realizowana w łączu komunikacyjnym typu punkt-punkt dla uproszczenia. W tym kontekście wszystkie podnośne OFDM są używane do multipleksowania danych przeznaczonych dla pojedynczego użytkownika w transmisji DL, a pojedynczy użytkownik jest przypisywany do wszystkich podnośnych w transmisji UL. Ze względu na niezależność między podnośnymi, transmisja OFDM może być również używana jako multipleksowanie użytkowników lub schemat wielokrotnej dostępności, umożliwiając jednoczesne transmisje z rozdzieloną częstotliwością z wieloma użytkownikami. W DL, podzbiór podnośnych OFDM jest używany do transmisji do jednego użytkownika, podczas gdy inny podzbiór podnośnych OFDM jest używany dla innego użytkownika. Podobnie, w UL, jeden użytkownik przesyła swoje dane przez podzbiór podnośnych OFDM, podczas gdy inny użytkownik może jednocześnie przesyłać swoje dane przez inny podzbiór podnośnych OFDM. Jak pokazano na rysunku 10.9, trzem rozproszonym przestrzennie użytkownikom przypisano trzy różne części podnośnych OFDM. Najprostszym sposobem

jest przydzielenie grupy kolejnych podnośnych do transmisji i odbioru użytkownika. Tymczasem podnośne dla użytkownika mogą być rozłożone w całym paśmie, aby wykorzystać różnorodność częstotliwości, za cenę nieco wyższej złożoności implementacji i wyższej podatności na uszkodzenia sprzętowe. Gdy OFDMA jest stosowany jako schemat wielokrotnego dostępu UL, przesyłane sygnały z różnych terminali muszą dotrzeć do stacji bazowej mniej więcej w tym samym czasie. Dokładniej rzecz biorąc, różnica czasu przybycia powinna być mniejsza niż długość prefiksu cyklicznego, aby zachować ortogonalność podnośnych i tym samym uniknąć zakłóceń międzynośnych między różnymi użytkownikami. Ze względu na różnicę opóźnień propagacji konieczne jest kontrolowanie czasu transmisji UL każdego terminala, np. pozwalając terminalowi znajdującemu się daleko od stacji bazowej wysłać swój sygnał z wyprzedzeniem. Taka kontrola czasu transmisji powinna regulować czas transmisji każdego terminala, aby zapewnić, że transmisje UL docierają do stacji bazowej mniej więcej w tym samym czasie. Ponadto, ponieważ czas propagacji zmienia się, gdy terminale przemieszczają się w komórce, kontrola czasu transmisji powinna być procesem dynamicznym, stale regulującym dokładny czas każdego terminala. Ponadto, istnieją zakłócenia międzynośnymi nawet w przypadku idealnej kontroli czasu transmisji z powodu błędów częstotliwości. Takie zakłócenia są zazwyczaj niskie przy rozsądnych błędach częstotliwości i rozproszeniu Dopplera. Zakłada się jednak, że różne podnośne są odbierane z co najmniej w przybliżeniu tym samym poziomem mocy. Odległości propagacji i odpowiadające im straty ścieżki mogą się znacznie różnić w UL. Stąd też siły odbieranego sygnału mogą się znacznie różnić, co oznacza potencjalnie znaczące zakłócenia z silniejszej podnośnej do jej słabszej sąsiedniej podnośnej, chyba że ortogonalność podnośnej jest idealnie zachowana. Aby tego uniknąć, może zaistnieć konieczność zastosowania w OFDMA przynajmniej pewnego stopnia kontroli mocy UL, co pozwoli na zmniejszenie mocy nadawczej terminali użytkowników znajdujących się w pobliżu stacji bazowej i zapewni, że wszystkie odbierane sygnały będą miały mniej więcej taki sam poziom mocy.

Dostęp wielokrotny z podziałem częstotliwości pojedynczej nośnej

Głównym wyzwaniem technicznym transmisji OFDM, podobnie jak każdej modulacji wielonośnej, są znaczne wahania chwilowej mocy przesyłanych sygnałów. Takie wahania mocy są zazwyczaj mierzone przez współczynnik mocy szczytowej do średniej (PAPR), który jest definiowany przez

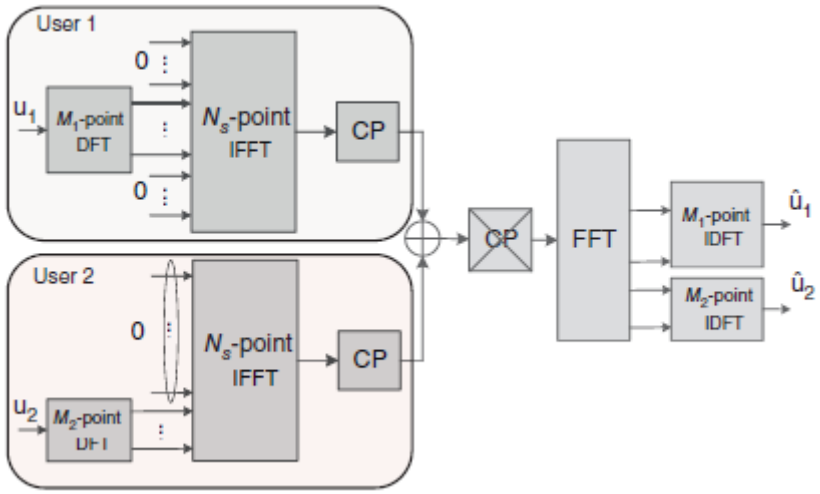
$$\mathrm{PAPR} := \frac{\max_{n=0,\ldots,N_s} \left(|s_m[n]|^2\right)}{\mathbb{E}\left[|s_m[n]|^2\right]}, \quad (91)$$

lub

$$\mathrm{PAPR} := \frac{\max_{mT \leq t < (m+1)T} \left(|s_m(t)|^2\right)}{\frac{1}{T}\int_{t=mT}^{(m+1)T} |s_m(t)|^2 dt} \quad (92)$$

Wysoki PAPR oznacza zmniejszoną wydajność wzmacniacza mocy i wyższy koszt wzmacniacza mocy. Jest to szczególne ograniczenie projektowe dla transmisji UL ze względu na wymagania niskiego zużycia energii i niskiego kosztu dla terminali mobilnych. Zaproponowano kilka metod, takich jak rezerwacja tonów, w której podzbiór podnośnych OFDM nie jest używany do transmisji danych, a zamiast tego jest modulowany w celu stłumienia największego szczytu, oraz selektywne szyfrowanie, które wybiera transmitowany sygnał o najniższym PAPR z szeregu zaszyfrowanych sygnałów przy użyciu różnych kodów. Jednak większość z tych metod ma ograniczenia, w jakim stopniu mogą obniżyć wahania mocy. Dlatego też atrakcyjne jest rozważenie szerokopasmowej transmisji pojedynczej nośnej, wykazującej stałą obwiednię z bardzo niskim PAPR, jako alternatywy dla transmisji wielonośnej, szczególnie w UL dla terminali mobilnych. Właściwość pojedynczej nośnej jest realizowana przez schemat transmisji

zwany DFT-spread OFDM lub DFT-s-OFDM, który ma niewielkie wahania w chwilowej mocy transmitowanych sygnałów i umożliwia elastyczne przydzielanie szerokości pasma. Podstawową zasadą DFT-s-OFDM jest wykonywanie wstępnego kodowania opartego na DFT w normalnej transmisji OFDM. Blok M symboli niosących informacje jest najpierw stosowany do DFT o rozmiarze M. Jego wyjście jest następnie przypisywane do podzbioru M kolejnych lub rozproszonych podnośnych modulatora OFDM implementowanego przez IFFT o rozmiarze Ns (ogólnie przyjmuje się, że N_s jest ustawione na potęgę 2, podczas gdy M jest bardziej elastyczne). Jeśli rozmiar DFT M jest równy rozmiarowi IFFT N_s, kaskadowe przetwarzanie DFT-IFFT będzie się wzajemnie wyrównywać, a przesyłany sygnał należy do transmisji pojedynczej nośnej. Jednak jeśli M jest mniejsze niż N_s, a pozostałe wejścia do IFFT są ustawione na zero, wyjście modulacji OFDM będzie sygnałem o niskich wahaniach mocy, wykazującym właściwość pojedynczej nośnej. Główną zaletą DFT-s-OFDM w porównaniu ze zwykłą transmisją OFDM jest niski PAPR w chwilowej mocy transmisji, co skutkuje zwiększoną wydajnością wzmacniacza mocy, co może obniżyć zużycie energii i umożliwić tanie terminale mobilne. Nominalna szerokość pasma przesyłanego sygnału dla użytkownika wynosi $M \Delta f$. Tak więc, poprzez dynamiczną regulację rozmiaru bloku M, chwilowa szerokość pasma przesyłanego sygnału może być zmieniana, co umożliwia elastyczne przydzielanie szerokości pasma. Ponadto, poprzez przypisanie wyjścia DFT do różnych podzbiorów podnośnych OFDM, przesyłany sygnał może być przesuwany w dziedzinie częstotliwości. Wielu użytkowników może jednocześnie przesyłać swoje dane, korzystając z transmisji DFT-s-OFDM, umożliwiając nie tylko zmiany niskiej mocy, takie jak transmisja pojedynczej nośnej, ale także OFDMA. Dlatego ta technika jest określana jako Single-Carrier Frequency-Division Multiple Access (SC-FDMA), która została przyjęta jako schemat transmisji UL w 3GPP Long-Term Evolution (LTE), a także 5G NR. Bez utraty ogólności, jak pokazano na rysunku , zakładamy dwóch użytkowników przesyłających



$$\mathbf{u}_1 = \left[u_1[0], u_1[1], \ldots, u_1[M_1 - 1]\right]^T \quad (93)$$

i

$$\mathbf{u}_2 = \left[u_2[0], u_2[1], \ldots, u_2[M_2 - 1]\right]^T, \quad (94)$$

odpowiednio w kierunku stacji bazowej. Każdy użytkownik wykonuje prekodowanie DFT na swoich przesłanych blokach symboli przed modulacją OFDM, np.

$$\tilde{u}_1[n] = \sum_{m=0}^{M_1-1} u_1[m] e^{-2\pi jnm/M_1}, \quad n = 0, 1, \ldots, M_1 - 1 \tag{95}$$

dla użytkownika 1. Następnie wstępnie zakodowane symbole dla użytkowników 1 i 2 są przypisywane do różnych części podnośnych OFDM, co skutkuje

$$\mathbf{a}_1 = \big[\underbrace{0, \ldots, 0}_{N_s - M_1 - M_2}, \tilde{u}_1[0], \ldots, \tilde{u}_1[M_1 - 1], \underbrace{0, \ldots, 0}_{M_2}\big]^T \tag{96}$$

i

$$\mathbf{a}_2 = \big[\underbrace{0, \ldots, 0}_{N_s - M_2}, \tilde{u}_2[0], \ldots, \tilde{u}_2[M_2 - 1]\big]^T \tag{97}$$

Pomijając dla uproszczenia osłabienia kanału i szum addytywny, odbiornik obserwuje połączony sygnał od użytkowników 1 i 2. Wykonaj demodulację FFT w punkcie N_s, a otrzymasz

$$\mathbf{a}_{rx} = \big[\underbrace{0, \ldots, 0}_{N_s - M_1 - M_2}, \tilde{u}_1[0], \ldots, \tilde{u}_1[M_1 - 1], \tilde{u}_2[0], \ldots, \tilde{u}_2[M_2 - 1]\big]^T \tag{98}$$

Oddzielenie symboli zależnych od użytkownika, a następnie utworzenie odpowiednio rozmiaru M1 i M2 IDFT, powoduje, że symbole informacyjne od dwóch użytkowników, tj. u_1 i u_2, są pomyślnie dostarczane.

Różnorodność opóźnień cyklicznych

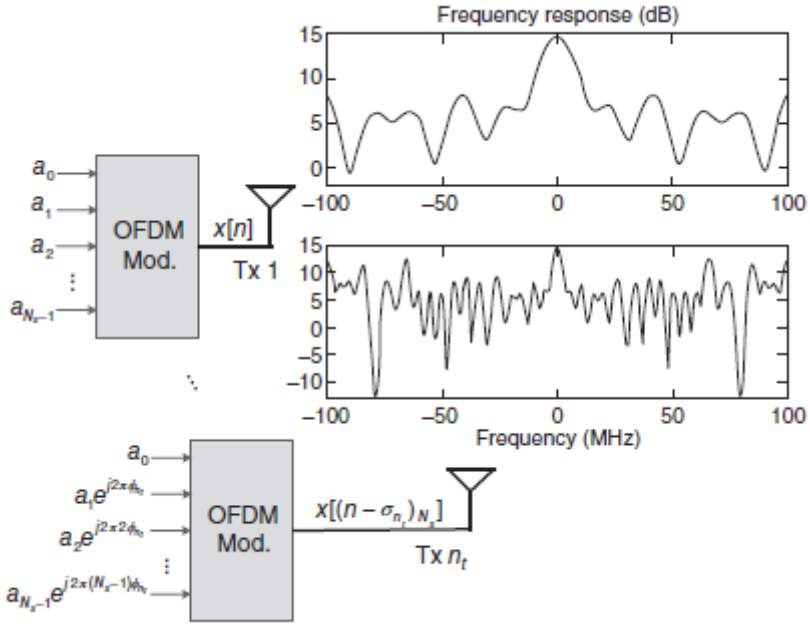
W szerokopasmowej transmisji jednonośnej, takiej jak szerokopasmowa CDMA, każdy symbol modulacji jest rozłożony na całej szerokości pasma sygnału. Ponieważ kanał jest wysoce selektywny częstotliwościowo, przesyłany sygnał doświadcza części częstotliwości o stosunkowo wysokich wzmocnieniach i części częstotliwości o wysokim tłumieniu. Taka transmisja informacji przez wiele części częstotliwości o różnej natychmiastowej jakości kanału uzyskuje zysk różnorodności częstotliwości. Z drugiej strony, w przypadku transmisji OFDM, każdy symbol modulacji jest ograniczony do wąskopasmowej podnośnej. Tak więc niektóre symbole modulacji mogą być całkowicie ograniczone do części częstotliwości o niskim natychmiastowym wzmocnieniu kanału. Dlatego też poszczególne symbole modulacji nie mogą wykorzystać różnorodności częstotliwości, nawet jeśli kanał jest wysoce selektywny częstotliwościowo w całej szerokości pasma transmisji OFDM. W konsekwencji prawdopodobieństwo błędu transmisji OFDM jest niskie i znacznie gorsze niż współczynnik błędu w szerokopasmowym systemie jednonośnym. Kanał radiowy podlegający dyspersji czasowej, z transmitowanym sygnałem propagującym się do odbiornika przez wiele niezależnie zanikających ścieżek z różnymi opóźnieniami, zapewnia możliwość różnorodności wielościeżkowej lub, co jest równoważne, różnorodności częstotliwości. Przypomnijmy, że odpowiedź impulsowa dla zanikającego kanału wielościeżkowego wynosi

$$h(\tau, t) = \sum_l a_l(t)\delta(\tau - \tau_l(t)), \tag{99}$$

gdzie $a_l(t)$ i $\tau_l(t)$ oznaczają odpowiednio zmienne w czasie tłumienie i opóźnienie propagacji ścieżki l. Jej odpowiedź częstotliwościowa jest obliczana przez Tse i Viswanatha

$$H(f;t) = \int_{-\infty}^{\infty} h(\tau,t)e^{-2\pi jf\tau}d\tau = \sum_{l} a_l(t)e^{-2\pi jf\tau_l(t)} \quad (100)$$

Istnieje różnicowa faza $2\pi f[\tau_{l1}(t) - \tau_{l2}(t)]$ pomiędzy różnymi ścieżkami, powodująca selektywne zanikanie częstotliwości. Oznacza to, że odpowiedź częstotliwościowa zmienia się znacząco, gdy f zmienia się o $B_c = 1/2T_d$, gdzie B_c jest szerokością pasma koherencji, a T_d oznacza rozproszenie opóźnienia. Oznacza to, że większe rozproszenie opóźnienia odpowiada szybszej zmianie odpowiedzi częstotliwościowej lub, co za tym idzie, większej selektywności częstotliwościowej. Jeśli sam kanał nie jest dyspersyjny w czasie lub selektywność częstotliwościowa jest niewystarczająca, można zastosować technikę zwaną różnorodnością opóźnień, aby stworzyć sztuczną dyspersję czasową lub, co za tym idzie, sztuczną selektywność częstotliwościową poprzez przesyłanie identycznych sygnałów z opóźnieniami przez wiele anten nadawczych. Indukowane opóźnienie powinno zostać określone w celu zapewnienia odpowiedniego poziomu selektywności częstotliwościowej w paśmie sygnału. Różnorodność opóźnień jest przejrzysta dla strony terminala, która obserwuje pojedynczy kanał radiowy podlegający dodatkowej dyspersji czasowej. Różnorodność opóźnień może być zatem bezpośrednio wykorzystana w istniejącym systemie komunikacji mobilnej bez żadnych problemów ze zgodnością w starszym standardzie interfejsu powietrznego. Różnorodność opóźnień cyklicznych (CDD) jest podobna do różnorodności opóźnień z tą główną różnicą, że działa blokowo i stosuje przesunięcia cykliczne, a nie opóźnienia liniowe, do różnych anten. Tak więc różnorodność opóźnień cyklicznych jest stosowana do schematów transmisji opartych na blokach, takich jak OFDM i DFT-s-OFDM. W przypadku transmisji OFDM cykliczne przesunięcie sygnału w dziedzinie czasu odpowiada zależnemu od częstotliwości przesunięciu fazowemu przed modulacją OFDM, jak pokazano na rysunku.



Podobnie jak różnorodność opóźnień, stworzy to sztuczną selektywność częstotliwości widzianą przez odbiornik. Aby uniknąć ograniczenia długości opóźnienia, CDD kołowo przesuwa próbki w symbolu OFDM zamiast dodawać liniowe opóźnienie do całego symbolu. Załóżmy, że nadajnik jest wyposażony w antenę składającą się z N_t elementów, indeksowanych przez n_t = 1, 2,…,Nt, co odpowiada cyklicznemu opóźnieniu σnt, odbiornik ma pojedynczą antenę. Zgodnie z twierdzeniem o przesunięciu DFT , kołowe przesunięcie sekwencji o skończonej długości odpowiada mnożeniu współczynnika fazy

w dziedzinie częstotliwości, a ta faza liniowo wzrasta wraz z indeksem. Odwołując się do równania (53), mamy oryginalną modulację OFDM

$$s[k] = \sum_{n=0}^{N_s-1} a_n e^{2\pi jnk/N_s} \quad (101)$$

a jego przesunięcie kołowe jest podane przez

$$\begin{aligned} s\left[(k-\sigma_{n_t})_{N_s}\right] &= \sum_{n=0}^{N_s-1} a_n e^{2\pi jn(k-\sigma_{n_t})/N_s} \\ &= \sum_{n=0}^{N_s-1} a_n e^{2\pi jnk/N_s} e^{-2\pi jn\sigma_{n_t}/N_s} \\ &= \sum_{n=0}^{N_s-1} \left(a_n e^{jn\phi_{n_t}}\right) e^{2\pi jnk/N_s} \end{aligned} \quad (102)$$

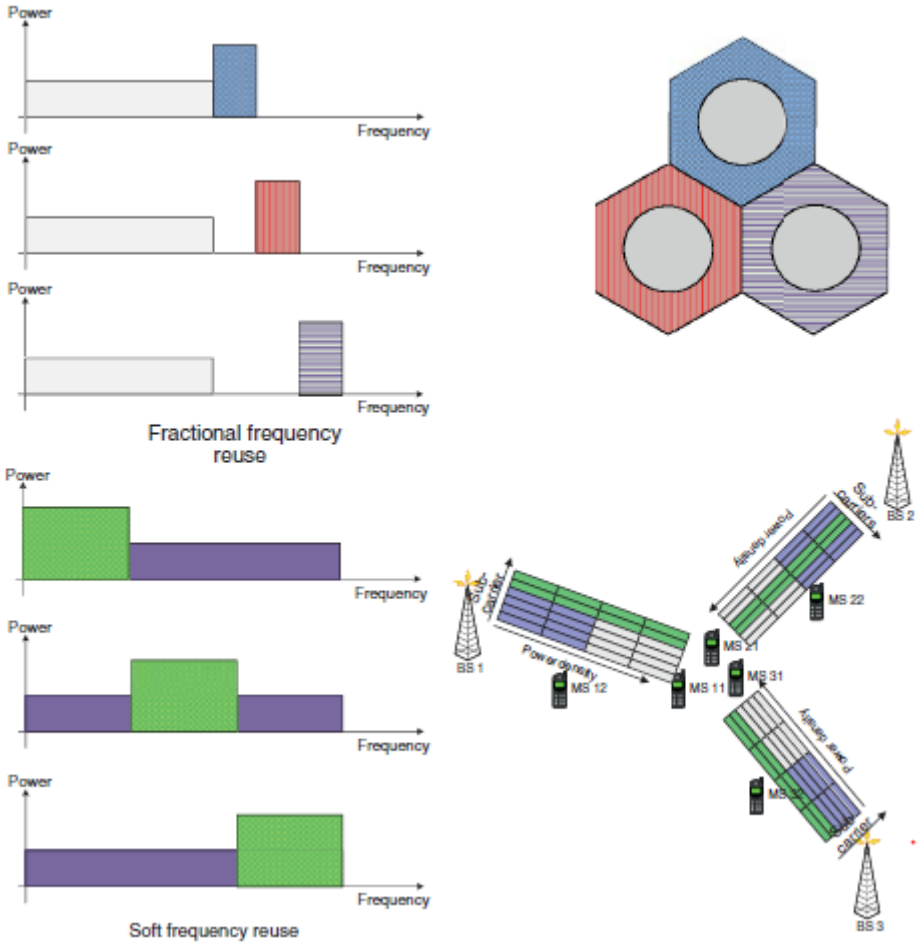
z $\phi_{n_t} = -\frac{2\pi j\sigma_{n_t}}{N_s}$.

Podsumowując, rozproszenie opóźnienia spowodowane propagacją wielodrogową w kanale bezprzewodowym zwiększa selektywność częstotliwości. Poprzez sztuczne indukowanie większego opóźnienia poprzez użycie wielu anten transmisyjnych, różnorodność opóźnień może zwiększyć szybkość zmian odpowiedzi częstotliwościowej kanału, ułatwiając wykorzystanie różnorodności częstotliwości. CDD jest szczególną implementacją różnorodności opóźnień w systemie OFDM, która zastępuje liniowe opóźnienie przesunięciem kołowym. Ponieważ obrót fazy w dziedzinie częstotliwości odpowiada cyklicznemu przesunięciu sygnału w dziedzinie czasu. Dodanie dodatkowego przesunięcia fazy do każdego przesyłanego symbolu przed modulacją DFT jako $\tilde{a}_n = a_n e^{jn\phi_{n_t}}$, gdzie ta faza liniowo wzrasta w odniesieniu do indeksu podnośnych, wygenerowana sekwencja OFDM staje się wersją przesuniętą kołowo oryginalnej sekwencji z przesunięciem σ_{nt}. Następnie odbiornik traktuje te sygnały tak samo jak komponenty wielodrogowe z pojedynczego nadajnika, z wyjątkiem większego rozproszenia opóźnienia, co zwiększa selektywność częstotliwości.

Wielokomórkowa OFDMA

We wczesnych systemach mobilnych stacja bazowa zazwyczaj pokrywała szeroki obszar o średnicy kilkudziesięciu kilometrów, montując antenę na dużej wysokości i transmitując sygnały radiowe o dużej mocy. Wszyscy użytkownicy telefonii komórkowej w tym obszarze zasięgu współdzielą przydzielone widmo, co prowadziło do ograniczonej pojemności. Aby sprostać rosnącemu zapotrzebowaniu na dużą pojemność, pojawiła się koncepcja sieci komórkowych poprzez ponowne wykorzystanie częstotliwości. W 1947 roku William R. Young po raz pierwszy przedstawił swój pomysł na heksagonalny układ komórek na szerokim obszarze, tak aby każdy telefon komórkowy mógł połączyć się z co najmniej jedną komórką. Następnie Douglas H. Ring rozwinął koncepcję Younga i naszkicował podstawowy projekt standardowej sieci komórkowej jako notatkę techniczną zatytułowaną Mobile Telephony - Wide Area Coverage z 11 grudnia 1947 roku . Ponieważ moc sygnału drastycznie spada wraz z odległością propagacji, to samo widmo częstotliwości można ponownie wykorzystać w lokalizacjach oddzielonych przestrzennie. Jeśli odległość jest wystarczająco duża, interferencja współkanałowa nie jest nie do przyjęcia. Podziel dostępne widmo na β nienakładające się na siebie części, a każda komórka w klastrze β sąsiadujących komórek jest przypisana do innej części. Dlatego ten sam kanał nie jest używany w

sąsiadujących komórkach, aby zmniejszyć zakłócenia współkanałowe. Współczynnik β oznacza, jak często kanał może być ponownie używany i jest nazywany współczynnikiem ponownego użycia częstotliwości. Poprzez ponowne użycie częstotliwości każda sąsiadująca komórka β, znana również jako klaster, dzieli całe widmo. W zależności od geometrii układu komórkowego i wzoru unikania zakłóceń współczynnik ponownego użycia może być różny. Klasyczny schemat unikania zakłóceń dzieli pasmo częstotliwości na np. trzy równe podpasma przydzielone komórkom, tak aby sąsiadujące komórki zawsze używały różnych częstotliwości. Ten schemat nazywa się twardym ponownym użyciem częstotliwości i prowadzi do niskich zakłóceń sąsiadujących komórek, z ceną dużej utraty pojemności, ponieważ tylko jedna trzecia zasobów jest używana w każdej komórce. Na przykład, dobrze znany standard GSM, który opiera się na technice wielodostępu TDMA, przyjął współczynnik ponownego wykorzystania wynoszący 3. Najprostszym schematem przydzielania częstotliwości w sieci komórkowej jest użycie współczynnika ponownego wykorzystania wynoszącego 1, tj. przydzielenie wszystkich fragmentów do każdej komórki, maksymalizując tilizację widma. Na przykład, system CDMA może użyć współczynnika ponownego wykorzystania wynoszącego 1, określanego jako uniwersalne ponowne wykorzystanie częstotliwości, z pomocą zaawansowanych technik tłumienia zakłóceń. Jednak w tym przypadku obserwuje się wysokie zakłócenia międzykomórkowe, szczególnie na krawędziach komórek, gdzie pożądany sygnał jest na najsłabszym poziomie, podczas gdy zakłócenia są na najsilniejszym poziomie. Ze względu na elastyczność transmisji wielonośnej, ponowne wykorzystanie częstotliwości można osiągnąć na poziomie granularności jednej podnośnej. Oznacza to, że różne podnośne można przypisać do różnych komórek, oprócz tradycyjnej metody ponownego wykorzystania częstotliwości, która może przydzielać różne nośne tylko w różnych komórkach. Aby zapewnić skuteczną Koordynację Inter-Cell Interference Coordination (ICIC) i maksymalizację wykorzystania widma, projektowanie systemów komórkowych opartych na OFDMA powinno poważnie uwzględniać schematy przydziału częstotliwości w oparciu o wiele komórek. Fractional Frequency Reuse (FFR) i Soft Frequency Reuse (SFR) to dwie szeroko uznawane metody, które zostały zaproponowane w celu poprawy efektywności widma i zmniejszenia ICI w systemie komunikacji wielonośnej. Zarówno w FFR, jak i SFR podnośne są podzielone na różne grupy, które są traktowane inaczej pod względem środka komórki i krawędzi komórki. W schemacie FFR szerokość pasma jest podzielona na dwa podpasma, a każda komórka jest odpowiednio podzielona na część wewnętrzną i część zewnętrzną. Jedno podpasmo jest dedykowane części wewnętrznej i ponownie wykorzystywane we wszystkich centrach komórkowych. Drugie podpasmo jest dalej podzielone na trzy nienakładające się na siebie części, które są przypisane do trzech sąsiadujących komórek, tak aby interferencja międzykomórkowa była minimalizowana na krawędzi komórki. Schemat SFR, zaproponowany przez Yanga w propozycji technicznej 3GPP R1-050507, wykorzystuje współczynnik ponownego wykorzystania częstotliwości wynoszący 1 w części wewnętrznej komórki i współczynnik ponownego wykorzystania częstotliwości wynoszący 3 w zewnętrznym obszarze komórki blisko krawędzi komórki. W części wewnętrznej komórki, poprzez ograniczenie mocy transmisji, powstają pewne odizolowane wyspy, które nie zakłócają się wzajemnie. Jak pokazano na Rysunku



, stacje mobilne 11 i 12 są połączone ze stacją bazową 1, stacje mobilne 21 i 22 są połączone ze stacją bazową 2, a stacje mobilne 31 i 32 są połączone ze stacją bazową 3. Stacje mobilne 11, 21 i 31 są zlokalizowane na przecięciu 3 komórek, stacje mobilne 12, 22 i 32 znajdują się w wewnętrznej części swoich odpowiednich komórek. W przypadku stacji mobilnej na skraju komórki przydzielane są im różne podnośne, aby uniknąć zakłóceń międzykomórkowych. W przypadku stacji mobilnych w pobliżu stacji bazowej dostępne są wszystkie podnośne, w porównaniu z jedynie ułamkiem wszystkich podnośnych w FFR. Gdy stosunek mocy między podnośnymi w centrum komórki a podnośnymi na skraju komórki wynosi 0, SFR jest równoważny twardemu ponownemu wykorzystaniu częstotliwości ze współczynnikiem 3. Gdy stosunek mocy osiąga 1, SFR jest równoważny uniwersalnemu ponownemu wykorzystaniu częstotliwości ze współczynnikiem 1. Poprzez dostosowanie stosunku mocy w zakresie od 0 do 1, można wdrożyć współczynnik ponownego wykorzystania częstotliwości z 3 do 1. Z tego powodu można go nazwać miękkim schematem ponownego wykorzystania częstotliwości.

Bezkomórkowe masowe MIMO-OFDMA

Większość zatorów ruchu w sieciach komórkowych występuje obecnie na skraju komórki. Tak zwane 95% prawdopodobnych szybkości transmisji danych użytkownika, które można zagwarantować 95% użytkowników i w ten sposób określić wydajność odczuwaną przez użytkownika, pozostają przeciętne w sieciach 5G. Rozwiązaniem tych problemów może być połączenie każdego użytkownika z wieloma rozproszonymi antenami. Jeśli w sieci jest tylko jedna ogromna komórka, nie pojawiają się żadne zakłócenia międzykomórkowe. Zaproponowano rozproszony system MIMO (ang. distributed massive multiple-input multiple-output), w którym duża liczba anten usługowych obsługuje znacznie mniejszą

liczbę autonomicznych użytkowników rozproszonych na dużym obszarze. Wszystkie anteny współpracują spójnie fazowo za pośrednictwem sieci fronthaul i obsługują wszystkich użytkowników w tym samym zasobie czasowo-częstotliwościowym. Nie ma komórek ani granic komórek. Dlatego system ten jest określany jako bezkomórkowe massive MIMO. Dzięki możliwości wykorzystania różnorodności przestrzennej przeciwko zanikaniu cienia bardziej efektywnie, rozproszony system może oferować znacznie większe prawdopodobieństwo zasięgu niż system kolokacyjny kosztem zwiększonych wymagań backhaul. Z drugiej strony, pasmo sygnału komunikacji mobilnej staje się coraz szersze, aby sprostać zapotrzebowaniu na większą przepustowość, szczególnie w pasmach wysokiej częstotliwości, takich jak komunikacja milimetrowa i terahercowa. Niemniej jednak komunikacja szerokopasmowa cierpi z powodu wysokiej selektywności częstotliwości zanikających kanałów. W związku z tym połączenie bezkomórkowego masywnego MIMO z transmisją OFDM, nazwanego bezkomórkowego masywnego MIMO-OFDM lub bezkomórkowego masywnego MIMO-OFDMA jest obiecujące w nadchodzących systemach 6G.

Model systemu

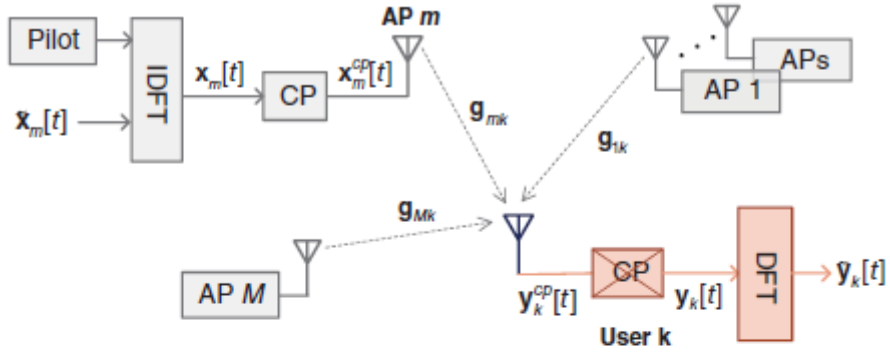
Rozważmy obszar geograficzny, w którym M losowo rozproszonych punktów dostępowych (AP) jest połączonych z jednostką centralną (CPU) za pośrednictwem sieci fronthaul i obsługuje K użytkowników. Bez utraty generacji załóżmy, że każdy AP i użytkownik jest wyposażony w jedną antenę, ale jego adaptacja do wieloantenowych AP jest prosta. W przeciwieństwie do konwencjonalnego Cell-Free massive MIMO (CFmMIMO), który wymaga $K \ll M$, liczba użytkowników w Cell-Free massive MIMO-OFDM (CFmMIMO-OFDM) jest skalowalna, od małego $K \ll M$ do bardzo dużego $K \gg M$. Użytkownicy są podzieleni na grupy, a każda grupa jest przypisana do różnych RB. Tak więc ograniczenie, że liczba użytkowników jest znacznie mniejsza niż liczba AP, jest nadal spełnione na każdej podnośnej lub RB. W transmisji CFmMIMO zakłada się, że zanikanie na małą skalę jest płaskie częstotliwościowo, modelowane przez kołowo symetryczną zmienną losową Gaussa o zerowej średniej i wariancji jednostkowej, tj. $h[t] \sim \mathcal{CN}(0,1)$. To założenie jest ważne tylko w przypadku komunikacji wąskopasmowej. Niemniej jednak większość obecnej i przyszłej komunikacji mobilnej jest szerokopasmowa i cierpi na poważną selektywność częstotliwościową. Kanał zanikania selektywny pod względem częstotliwości jest modelowany jako filtr liniowy o zmiennym czasie $h[t] = [\, h_0[t],\ldots, h_{L-1}[t] \,]^T$, gdzie długość filtra L jest związana z rozproszeniem opóźnienia wielościeżkowego T_d i interwałem próbkowania T_s. Wzmocnienie odczepu jest obliczane przez

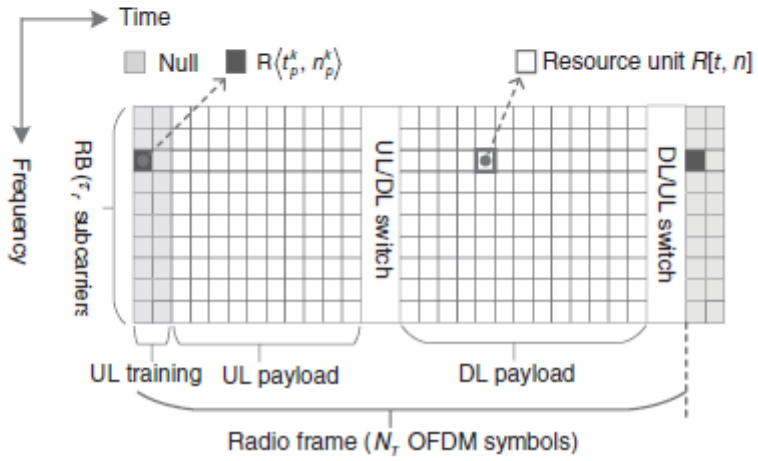
$$h_l[t] = \sum_i a_i(tT_s) e^{-j2\pi f_c \tau_i(tT_s)} \operatorname{sinc}\left[l - \frac{\tau_i(tT_s)}{T_s}\right] \tag{103}$$

dla $l = 0,\ldots, L-1$, z częstotliwością nośną f_c i zmiennym w czasie tłumieniem $a_i(t)$ i opóźnieniem $\tau_i(t)$ i-tej ścieżki sygnału. Kanał zanikania między AP m i użytkownikiem k jest podany przez

$$\begin{aligned} \mathbf{g}_{mk}[t] &= \left[g_{mk,0}[t], \ldots, g_{mk,L_{mk}-1}[t]\right]^T \\ &= \sqrt{\beta_{mk}[t]}\left[h_{mk,0}[t], \ldots, h_{mk,L_{mk}-1}[t]\right]^T = \sqrt{\beta_{mk}[t]}\mathbf{h}_{mk}[t], \end{aligned} \tag{104}$$

gdzie $g_{mk,l}[t] = \sqrt{\beta_{mk}[t]}h_{mk,l}[t]$ i $\beta_{mk}[t]$ oznacza zanikanie na dużą skalę, które jest niezależne od częstotliwości i zmienia się powoli, a Lmk oznacza długość kanału. Transmisja danych w systemie OFDM jest zorganizowana blokowo, jak pokazano na rysunku.





Piszemy

$$\tilde{\mathbf{x}}_m[t] = \left[\tilde{x}_{m,0}[t], \ldots, \tilde{x}_{m,n}[t], \ldots, \tilde{x}_{m,N-1}[t]\right]^T \quad (105)$$

aby oznaczyć blok transmisji w dziedzinie częstotliwości AP m na t-tym symbolu OFDM. Przekształć $\tilde{\mathbf{x}}_m[t]$ w sekwencję w dziedzinie czasu

$$\mathbf{x}_m[t] = \left[x_{m,0}[t], \ldots, x_{m,k}[t], \ldots, x_{m,N-1}[t]\right]^T \quad (106)$$

poprzez N-punktową IDFT, tj.

$$x_{m,k}[t] = \frac{1}{N}\sum_{n=0}^{N-1} \tilde{x}_{m,n}[t] e^{2\pi jkn/N} \quad (107)$$

dla k = 0, 1,…,N − 1. Definiowanie macierzy DFT

$$\mathbf{F} = \begin{bmatrix} \omega_N^{0\cdot 0} & \cdots & \omega_N^{0\cdot(N-1)} \\ \vdots & \ddots & \vdots \\ \omega_N^{(N-1)\cdot 0} & \cdots & \omega_N^{(N-1)\cdot(N-1)} \end{bmatrix} \quad (108)$$

z pierwotnym pierwiastkiem N-tego stopnia z jedności $\omega^{n.k}_N = e^{2\pi jnk/N}$, modulację OFDM można zapisać w postaci macierzowej jako

$$\mathbf{x}_m[t] = \mathbf{F}^{-1}\bar{\mathbf{x}}_m[t] = \frac{1}{N}\mathbf{F}^*\bar{\mathbf{x}}_m[t]. \quad (109)$$

Prefiks cykliczny o długości N_{CP} jest dodawany pomiędzy dwoma kolejnymi blokami, aby uniknąć interferencji między symbolami i zachować ortogonalność podnośnych. W ten sposób transmitowany sygnał jest wyrażony przez

$$\mathbf{x}_m^{\mathrm{CP}}[t] = \left[\underbrace{x_{m,N-N_{\mathrm{CP}}}[t], \ldots, x_{m,N-1}[t]}_{\text{Cyclic prefix}}, x_{m,0}[t], \ldots, x_{m,N-1}[t]\right]^T \quad (110)$$

Sygnał $x^{CP}_m[t]$ przechodzi przez kanał $g_{mk}[t]$, aby dotrzeć do typowego użytkownika k, co skutkuje $x^{CP}_m[t] * g_{mk}[t]$, gdzie $*$ oznacza splot liniowy. Zatem całkowity odebrany sygnał u użytkownika k to $y^{CP}_k[t] = \Sigma^M_{m=1} x^{CP}_m[t] * g_{mk}[t] + z_k[t]$, gdzie $z_k[t]$ jest wektorem szumu addytywnego. Usunięcie CP daje

$$\mathbf{y}_k[t] = \sum_{m=1}^{M} \mathbf{g}_{mk}^N[t] \otimes \mathbf{x}_m[t] + \mathbf{z}_k[t]. \quad (111)$$

gdzie $\otimes$ oznacza splot cykliczny, a $g^N_{mk}[t]$ to filtr kanału N-punktowego utworzony przez dodanie zer na końcu $g_{mk}[t]$, tj.

$$\mathbf{g}_{mk}^N[t] = \left[g_{mk,0}[t], \ldots, g_{mk,L_{mk}-1}[t], 0, \ldots, 0\right]^T \quad (112)$$

Demodulator DFT wyprowadza odebrany sygnał w dziedzinie częstotliwości

$$\bar{\mathbf{y}}_k[t] = \mathbf{F}\mathbf{y}_k[t]. \quad (113)$$

Podstawiając równania (109) i (111) do równania (113) i stosując twierdzenie splotu dla DFT, wiemy, że

$$\begin{aligned}\bar{\mathbf{y}}_k[t] &= \sum_{m=1}^{M} \mathbf{F}\left(\mathbf{g}_{mk}^N[t] \otimes \mathbf{x}_m[t]\right) + \mathbf{F}\mathbf{z}_k[t] \\ &= \sum_{m=1}^{M} \bar{\mathbf{g}}_{mk}[t] \odot \bar{\mathbf{x}}_m[t] + \bar{\mathbf{z}}_k[t],\end{aligned} \quad (114)$$

gdzie $\odot$ oznacza iloczyn Hadamarda (mnożenie elementów), odpowiedź kanału w dziedzinie częstotliwości i szum są podane wzorem

$$\bar{\mathbf{g}}_{mk}[t] = \mathbf{F}\mathbf{g}_{mk}^N[t] \quad (115)$$

i

$$\bar{\mathbf{z}}_k[t] = \mathbf{F}\mathbf{z}_k[t], \quad (116)$$

odpowiednio. Na koniec kanał selektywny częstotliwościowo jest przekształcany w zestaw N niezależnych płaskich częstotliwościowo podnośnych. Transmisja sygnału w DL na n-tej podnośnej jest podana przez

$$\bar{y}_{k,n}[t] = \sum_{m=1}^{M} \bar{g}_{mk,n}[t]\bar{x}_{m,n}[t] + \bar{z}_{k,n}[t], \quad k \in \{1, \ldots, K\} \quad (117)$$

gdzie $\bar{g}_{mk,n}[t]$ jest n-tym elementem $\bar{g}_{mk,n}[t]$. Podobnie transmisja UL jest wyrażona przez

$$\bar{y}_{m,n}[t] = \sum_{k=1}^{K} \bar{g}_{mk,n}[t]\bar{x}_{k,n}[t] + \bar{z}_{m,n}[t], \quad m \in \{1, \ldots, M\} \quad (118)$$

Proces komunikacji

Transmisja DL z AP do użytkowników i UL od użytkowników do AP są rozdzielone przez dupleks z podziałem czasu (TDD) przy założeniu doskonałej wzajemności kanałów.Ramka radiowa jest głównie podzielona na trzy fazy: szkolenie UL, transmisja ładunku UL i transmisja ładunku DL.

Trening łącza w górę

Piszemy $\mathcal{R}\langle t, n\rangle$, aby oznaczyć jednostkę zasobów (RU) na n-tej podnośnej t-tego symbolu OFDM. Zasób czasowo-częstotliwościowy ramki radiowej jest podzielony na NRB RB, z których każdy zawiera λRB = N⁄NRB (przyjęte jako liczba całkowita) kolejnych podnośnych, jak pokazano na rysunku r-ty RB jest zdefiniowany jako

$$B_r \triangleq \{\mathcal{R}\langle t, n\rangle | 1 \le t \le N_T \text{ and } (r-1)\lambda_{\mathrm{RB}} \le n < r\lambda_{\mathrm{RB}}\}, \quad (119)$$

dla dowolnego r ∈ {1,…,N_{RB}}. Transmisja ramki radiowej w CFmMIMO odbywa się w czasie koherencji, a szerokość jednego RB jest mniejsza niż szerokość pasma koherencji. Korzystając z korelacji czasu i częstotliwości, współczynnik kanału dowolnego RU można uzyskać poprzez interpolację oszacowań kanału pilotów. Bez utraty ogólności przyjmuje się model zanikania bloku, w którym zakłada się, że współczynniki kanału dla wszystkich RU w jednym RB są identyczne, tj.

$$\bar{g}_{mk,n}[t] = \bar{g}^r_{mk} \iff \mathcal{R}\langle t, n\rangle \in B_r. \quad (120)$$

Oszacowanie kanału konwencjonalnego systemu CFmMIMO opiera się na sekwencjach pilota w dziedzinie czasu, gdzie maksymalna liczba sekwencji ortogonalnych wynosi τ_p przy użyciu τ_p symboli pilota. Jeśli K ⩽ τ_p, można uniknąć zanieczyszczenia pilota. Jednak ze względu na ograniczenie długości ramki niektórzy użytkownicy muszą współdzielić tę samą sekwencję, gdy K > τ_p, co prowadzi do zanieczyszczenia pilota. Natomiast CFmMIMO-OFDM jest w stanie zapewnić więcej ortogonalnych pilotów za pomocą multipleksowania z podziałem częstotliwości dzięki dodatkowemu stopniowi swobody uzyskanemu z dziedziny częstotliwości. Aby oszacować $\bar{g}^r_{mk}$, każdy użytkownik na B_r potrzebuje tylko jednego symbolu pilota. Załóżmy, że pierwsze τ_p symboli OFDM są dedykowane do treningu UL, jeden RB ma $N_p = \tau_p\lambda_{RB}$ ortogonalnych pilotów. Liczba użytkowników przydzielonych do B_r jest oznaczona przez K_r, nie ma zanieczyszczenia pilota, jeśli $K_r \leqslant N_p$, co jest bardzo rozluźnionym

warunkiem. Piszemy $\mathcal{R}\langle t^k_{pz} n^k_p\rangle$ z $1 \leqslant t^k_p \leqslant \tau_p$ i $(r-1)\lambda_{RB} \leqslant n^k_p < r\lambda_{RB}$, aby oznaczyć RU zarezerwowane dla symbolu pilota użytkownika k, $k \in \{1,\ldots, K_r\}$. Inni użytkownicy zachowują ciszę (null) w tym RU, aby osiągnąć ortogonalność. Matematycznie przypisanie pilota jest określone przez

$$\bar{x}_{k,n}[t] = \begin{cases} \sqrt{p_u}\,\mathbb{P}_k, & \text{if } t = t_p^k \wedge n = n_p^k \\ 0, & \text{otherwise} \end{cases}, \quad 1 \le t \le \tau_p. \tag{121}$$

gdzie $\wedge$ oznacza logiczne AND, $\mathbb{P}k$ jest znanym symbolem pilota z $\mathbb{E}[|\mathbb{P}_k|^2] = 1$, a p_u oznacza limit mocy transmisji UL. Podstawienie równań (120) i (121) do równania (118) daje odebrany sygnał m-tego AP na $\mathcal{R}\langle t^k_p, n^k_p\rangle$

$$\begin{aligned} \bar{y}_{m,n_p^k}[t_p^k] &= \sum_{k=1}^{K_r} \bar{g}_{mk,n_p^k}[t_p^k]\bar{x}_{k,n_p^k}[t_p^k] + \bar{z}_{m,n_p^k}[t_p^k] \\ &= \bar{g}_{mk,n_p^k}[t_p^k]\bar{x}_{k,n_p^k}[t_p^k] + \sum_{k'\neq k}^{K_r} \bar{g}_{mk',n_p^k}[t_p^k]\bar{x}_{k',n_p^k}[t_p^k] + \bar{z}_{m,n_p^k}[t_p^k] \\ &= \sqrt{p_u}\bar{g}_{mk,n_p^k}[t_p^k]\mathbb{P}_k + \bar{z}_{m,n_p^k}[t_p^k] \\ &= \sqrt{p_u}\bar{g}^r_{mk}\mathbb{P}_k + \bar{z}_{m,n_p^k}[t_p^k]. \end{aligned} \tag{122}$$

Niech $\hat{g}^r_{mk}$ będzie oszacowanie $\hat{g}^r_{mk}$, mamy $\hat{g}^r_{mk} = \bar{g}^r_{mk} - \xi^r_{mk}$ z błędem oszacowania ξ^r_{mk} podniesionym przez szum addytywny. Zastosowanie oszacowania MMSE daje

$$\hat{g}^r_{mk} = \left(\frac{R_{gg}\mathbb{P}_k^*}{R_{gg}|\mathbb{P}_k|^2 + R_{nn}}\right)\bar{y}_{m,n_p^k}[t_p^k] = \left(\frac{\beta_{mk}\mathbb{P}_k^*}{\beta_{mk}|\mathbb{P}_k|^2 + \sigma_z^2}\right)\bar{y}_{m,n_p^k}[t_p^k], \tag{123}$$

który dotyczy $R_{gg} = \mathbb{E}\left\{\left|\bar{g}^r_{mk}\right|^2\right\} = \beta_{mk}$ and $R_{nn} = \mathbb{E}\left\{\left|\bar{z}_{m,n}[t]\right|^2\right\} = \sigma_z^2$. Oblicz wariancję $\hat{g}^r_{mk}$ jako

$$\begin{aligned} \mathbb{E}\left[\bar{g}^r_{mk}(\hat{g}^r_{mk})^*\right] &= \mathbb{E}\left[\frac{\beta_{mk}^2|\mathbb{P}_k|^2\left|\sqrt{p_u}\bar{g}^r_{mk}\mathbb{P}_k + \bar{z}_{m,n_p^k}[t_p^k]\right|^2}{(\beta_{mk}|\mathbb{P}_k|^2 + \sigma_z^2)^2}\right] \\ &= \frac{\beta_{mk}^2\mathbb{E}\left[\left|\sqrt{p_u}\bar{g}^r_{mk}\mathbb{P}_k + \bar{z}_{m,n_p^k}[t_p^k]\right|^2\right]}{(\beta_{mk} + \sigma_z^2)^2} \\ &= \frac{p_u\beta_{mk}^2}{p_u\beta_{mk} + \sigma_z^2}. \end{aligned} \tag{124}$$

W konsekwencji wiemy, że $\hat{g}^r_{mk} \in \mathcal{CN}(0, \alpha_{mk})$ gdzie $\alpha_{mk} = \frac{p_u\beta_{mk}^2}{p_u\beta_{mk}+\sigma_z^2}$, w porównaniu z $\bar{g}^r_{mk} \in \mathcal{CN}(0, \beta_{mk})$.

Transmisja danych ładunku łącza w górę

Załóżmy, że symbole OFDM τ_u są używane do transmisji UL, w RU $\mathcal{R}\langle t, n\rangle \in B_r$, $\tau_p < t \le \tau_p + \tau_u$, wszyscy użytkownicy Kr jednocześnie przesyłają swoje sygnały do punktów dostępowych. Użytkownik k-ty waży swój symbol transmisji $q_{k,n}[t]$, spełniając

$$\mathbb{E}\left[|q_{k,n}[t]|^2\right] = 1. \quad (125)$$

za pomocą współczynnika kontroli potęgi $\sqrt{\psi_k}$, $0 \le \psi_k \le 1$. Podstawienie $\bar{x}_{k,n}[t] = \sqrt{\psi_k p_u} q_{k,n}[t]$ do równania (118) daje

$$\bar{y}_{m,n}[t] = \sqrt{p_u} \sum_{k=1}^{K_r} \bar{g}_{mk,n}[t] \sqrt{\psi_k} q_{k,n}[t] + \bar{z}_{m,n}[t]. \quad (126)$$

Transmisja danych ładunku łącza w dół

Podczas gdy CFmMIMO stosuje sprzężone formowanie wiązki w DL, CFmMIMO-OFDM wykorzystuje sprzężone formowanie wiązki w dziedzinie częstotliwości. W $\mathcal{R}\langle t, n\rangle \in B_r$, $\tau_p + \tau_u < t \le N_T$, każdy AP multipleksuje łącznie symbole K_r, tj. $s_{k,n}[t]$ przeznaczone dla użytkownika k, $k = 1,\ldots, K_r$, przed transmisją. Przy współczynniku sterowania mocą $\sqrt{\eta_{mk}}$, $0 \le \eta_{mk} \le 1$, przesyłany sygnał m-tego AP wynosi

$$\bar{x}_{m,n}[t] = \sqrt{p_d} \sum_{k=1}^{K_r} \sqrt{\eta_{mk}} \left(\hat{g}_{mk,n}[t]\right)^* s_{k,n}[t]. \quad (127)$$

Podstawiając równanie (127) do równania (117) w celu uzyskania sygnału odebranego dla użytkownika k

$$\begin{aligned}
\bar{y}_{k,n}[t] &= \sqrt{p_d} \sum_{m=1}^{M} \bar{g}_{mk,n}[t] \sum_{k'=1}^{K_r} \sqrt{\eta_{mk'}} \left(\hat{g}_{mk',n}[t]\right)^* s_{k',n}[t] + \bar{z}_{k,n}[t] \\
&= \underbrace{\sqrt{p_d} \sum_{m=1}^{M} \sqrt{\eta_{mk}} \left|\hat{g}_{mk,n}[t]\right|^2 s_{k,n}[t]}_{\text{Desired signal}} \\
&+ \underbrace{\sqrt{p_d} \sum_{m=1}^{M} \hat{g}_{mk,n}[t] \sum_{k' \ne k}^{K_r} \sqrt{\eta_{mk'}} \left(\hat{g}_{mk',n}[t]\right)^* s_{k',n}[t]}_{\text{Multi-user interference}} \\
&+ \underbrace{\sqrt{p_d} \sum_{m=1}^{M} \xi_{mk,n}[t] \sum_{k'=1}^{K_r} \sqrt{\eta_{mk'}} \left(\hat{g}_{mk',n}[t]\right)^* s_{k',n}[t]}_{\text{Channel-estimate error}} + \underbrace{\bar{z}_{k,n}[t]}_{\text{Noise}}.
\end{aligned} \quad (128)$$

Zakłada się, że każdy użytkownik ma wiedzę na temat statystyk kanału

$$\mathbb{E}\left[\sum_{m=1}^{M} \sqrt{\eta_{mk}} \left|\hat{g}_{mk,n}[t]\right|^2\right] \quad (129)$$

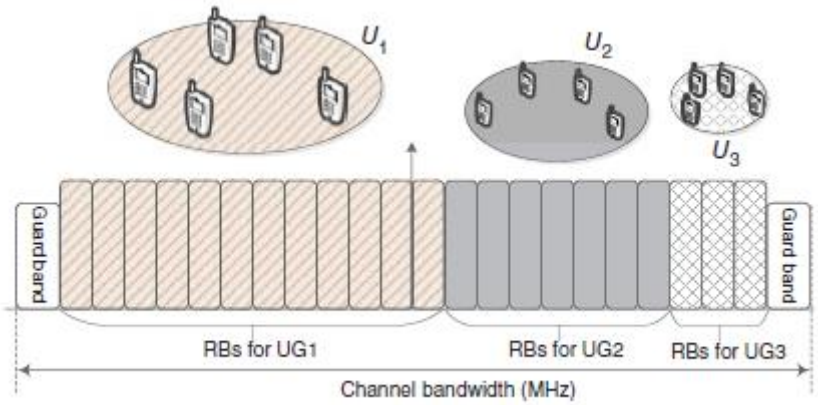
zamiast realizacji kanałowych $\hat{g}_{mk,n}[t]$, ponieważ w DL nie ma oszacowania pilota i kanału. Możemy wywnioskować, że wydajność widmowa użytkownika k on$\mathcal{R}\langle t, n\rangle \in B_r$ jest ograniczona dolną granicą $\log_2(1 + \gamma^{\langle t,n\rangle}_k)$ z

$$\gamma_k^{\langle t,n\rangle} = \frac{p_d\left(\sum_{m=1}^{M}\sqrt{\eta_{mk}}\alpha_{mk}\right)^2}{\sigma_z^2 + p_d\sum_{m=1}^{M}\beta_{mk}\sum_{k'=1}^{K_r}\eta_{mk'}\alpha_{mk'}} \quad (130)$$

co oznacza, że efekt zanikania na małą skalę zanika (znany również jako utwardzanie Hannela), tak że $\gamma^{\langle t,n\rangle}{}_k = \gamma^r{}_k$, dla wszystkich $\mathcal{R}\langle t, n\rangle \in B_r$.

Przydział zasobów specyficzny dla użytkownika

Konwencjonalne systemy CFmMIMO mogą obsługiwać tylko bardzo niewielu użytkowników K ≪ M przy zachowaniu jednolitej jakości usług. Wykorzystując domenę częstotliwości, CFmMIMO-OFDM dostosowuje się do różnej liczby użytkowników, od kilku K ≪ M do ogromnej liczby K ≫ M, i jest elastyczny, aby oferować różne szybkości transmisji danych dla heterogenicznych użytkowników. Jak pokazano na rysunku,



sklasyfikuj wszystkich użytkowników

$$\mathcal{U} = \{u_1, u_2, \ldots, u_K\} \quad (131)$$

na różne grupy $\mathcal{U}$s, s = 1, 2,…, S pod względem ich wymagań dotyczących przepustowości danych, poddając je

$$\bigcup_{s=1}^{S}\mathcal{U}_s = \mathcal{U} \quad (132)$$

i

$$\mathcal{U}_s \bigcap \mathcal{U}_{s'} = \emptyset, \quad \forall s' \neq s. \quad (133)$$

Liczba użytkowników w $\mathcal{U}_s$ spełnia$||\mathcal{U}_s|| \ll M$ i $\Sigma^S{}_{s=1}||\mathcal{U}s|| = K$, gdzie $|\cdot|$ oznacza kardynalność zbioru. Pula zasobów to $\mathbb{B} = \{B_r | 1 \le r \le N_{RB}\}$, gdzie granularność dla alokacji wynosi jeden RB. Używając $\mathbb{B}_s$ do oznaczenia RB przydzielonych do$\mathcal{U}$s, mamy $\bigcup^S{}_{s=1}\mathbb{B}_s \in \mathbb{B}$ (gdy używane są wszystkie RB, $\bigcup^S{}_{s=1}\mathbb{B}_s = \mathbb{B}$) i $\mathbb{B}_s \cap \mathbb{B}_{s'} = \emptyset\ \forall s' \neq s$. Jeżeli $B_r \in \mathbb{B}_s$, liczba użytkowników obsługiwanych przez tę RB wynosi $K_r = |\mathcal{U}_s|$. Dla dowolnego użytkownika $k \in \mathcal{U}_s$ jego szybkość transmisji danych na użytkownika w DL wynosi

$$R_k = \left(1 - \frac{\tau_p + \tau_u}{N_T}\right) \sum_{B_r \in \mathrm{B}_s} \lambda_{\mathrm{RB}} \triangle f \log_2 \left(1 + \gamma_k^r\right) \quad (134)$$

gdzie Δf jest odstępem między podnośnymi. Wówczas suma przepustowości danych DL systemu wynosi $R_d = \Sigma^S_{s=1} \Sigma_{k \in \mathcal{U}_s} R_k$.

Nieortogonalny wielokrotny dostęp

Projekt wydajnej techniki wielokrotnego dostępu jest krytycznym aspektem systemu komórkowego, który był dominującym sposobem rozróżniania różnych komunikacji bezprzewodowych od pierwszej do piątej generacji. Na przykład FDMA był używany w 1G, wielokrotny dostęp z podziałem czasu w większości 2G, a wielokrotny dostęp z podziałem kodu w 3G. W 4G OFDMA i SC-FDMA zostały przyjęte odpowiednio do transmisji DL i UL. Większość z tych technik opiera się na filozofii wielokrotnego dostępu ortogonalnego (OMA), w której ortogonalna jednostka zasobów, np. przedział czasowy, podnośna i ortogonalny kod rozproszony, jest dedykowana pojedynczemu użytkownikowi. W ten sposób zysk multipleksowania jest osiągany przy rozsądnej złożoności, jednocześnie łagodząc zakłócenia wielu użytkowników. Aby sprostać heterogenicznym wymaganiom dotyczącym łączności masowej, wysokiej wydajności widmowej, niskiego opóźnienia i zwiększonej uczciwości, system 5G przyjął zupełnie nową technikę zwaną Non-Orthogonal Multiple Access (NOMA) jako jedną ze swoich metod wielodostępu. W przeciwieństwie do konwencjonalnych schematów OMA, kluczową cechą wyróżniającą NOMA jest obsługa większej liczby użytkowników niż liczba jednostek zasobów ortogonalnych za pomocą nieortogonalnego współdzielenia zasobów, przy cenie wyrafinowanej eliminacji zakłóceń między użytkownikami w odbiorniku. Ding i inni podają prosty przykład ilustrujący wyższość NOMA w porównaniu z OMA. Rozważmy scenariusz, w którym użytkownik o bardzo złych warunkach kanału musi zostać obsłużony w celach uczciwości, np. ten użytkownik ma dane o wysokim priorytecie lub nie był obsługiwany przez długi czas. W tym przypadku użycie OMA oznacza, że nieuniknione jest, że jeden z ograniczonych zasobów pasma jest zajmowany wyłącznie przez tego użytkownika, pomimo jego złych warunków kanału. Taka konstrukcja szkodzi wydajności widmowej i pojemności całego systemu. W takiej sytuacji użycie NOMA zapewnia, że użytkownik ze słabymi warunkami kanału zostanie obsłużony, a użytkownicy z lepszymi warunkami kanału będą mogli jednocześnie uzyskać dostęp do tych samych zasobów. W związku z tym, jeśli ma być zagwarantowana sprawiedliwość użytkowników, pojemność systemu NOMA może być znacznie większa niż OMA. Oprócz wzrostu wydajności widmowej, NOMA może skutecznie obsługiwać więcej użytkowników, zapewniając masową łączność w scenariuszu wdrażania Internetu rzeczy.

Podstawy NOMA

Chociaż interferencja między użytkownikami wśród użytkowników ortogonalnie multipleksowanych umożliwia detekcję wielu użytkowników (MUD) o niskiej złożoności w odbiorniku, powszechnie wiadomo, że OMA nie może osiągnąć łącznej przepustowości systemu bezprzewodowego dla wielu użytkowników. Kodowanie superpozycji w nadajniku i sukcesywne usuwanie zakłóceń (SIC) w odbiorniku umożliwiają ponowne wykorzystanie każdej jednostki zasobów ortogonalnych przez więcej niż jednego użytkownika. Po stronie nadajnika wszystkie pojedyncze symbole informacyjne są nakładane na jedną falę, podczas gdy SIC po stronie odbiornika dekoduje sygnały iteracyjnie, aż otrzyma pożądany sygnał. Ten schemat jest czasami nazywany NOMA w domenie mocy. Podstawowa zasada NOMA i porównanie jej sumarycznej przepustowości z OMA zostaną przedstawione w tej sekcji odpowiednio w odniesieniu do transmisji DL i UL.

Nieortogonalne multipleksowanie łącza w dół

W DL stacja bazowa z pojedynczą anteną nakłada symbole niosące informacje przeznaczone dla użytkowników K z pojedynczą anteną. Zmultipleksowany sygnał w stacji bazowej można wyrazić za pomocą

$$s = \sum_{k=1}^{K} \sqrt{\alpha_k P_d} s_k, \quad (135)$$

gdzie s_k jest symbolem informacyjnym dla ogólnego użytkownika k, k = 1, 2,..., K, spełniającym $\mathbb{E}[|s_k|^2] = 1$, α_k reprezentuje współczynnik przydziału mocy podlegający $\Sigma^K_{k=1} \alpha_k \leqslant 1$, a P_d oznacza całkowitą moc nadawczą stacji bazowej. Wyzwaniem jest podjęcie decyzji, jak przydzielić moc użytkownikom, co jest krytyczne dla anulowania zakłóceń w odbiorniku. Dlatego NOMA jest uważane za rodzaj wielokrotnego dostępu w domenie mocy. Ogólnie rzecz biorąc, więcej mocy jest przydzielane użytkownikowi o mniejszym wzmocnieniu kanału, np. znajdującemu się dalej od stacji bazowej, w celu poprawy odbieranego przez niego SNR, tak aby można było zagwarantować wysoką niezawodność wykrywania. Pomimo mniejszej mocy przydzielonej użytkownikowi o silniejszym wzmocnieniu kanału, np. blisko stacji bazowej, jest on w stanie prawidłowo wykryć swój sygnał z rozsądnym SNR. Użytkownik i obserwuje odebrany sygnał

$$\begin{aligned} r_i &= g_i s + n_i \\ &= g_i \sum_{k=1}^{K} \sqrt{\alpha_k P_d} s_k + n_i \\ &= \underbrace{g_i s_i}_{\text{Desired signal}} + \underbrace{g_i \sum_{k=1, k\neq i}^{K} \sqrt{\alpha_k P_d} s_k}_{\text{Multi-user interference}} + n_i, \end{aligned} \quad (136)$$

gdzie g_k oznacza zespolone wzmocnienie kanału między stacją bazową a użytkownikiem k, ni to addytywny biały szum gaussowski o zerowej średniej i gęstości widmowej mocy N_0 (W/Hz). Bez utraty ogólności możemy założyć, że użytkownik 1 ma największe wzmocnienie kanału, a użytkownik K jest najsłabszy, tj.

$$|g_1| \geqslant |g_2| \geqslant \cdots \geqslant |g_K|, \quad (137)$$

Ten sam sygnał s, który zawiera wszystkie symbole informacyjne, jest dostarczany do wszystkich użytkowników. Optymalna kolejność usuwania zakłóceń to wykrywanie użytkownika z największym przydziałem mocy (najsłabszym wzmocnieniem kanału) do użytkownika z najmniejszym przydziałem mocy (najsilniejszym wzmocnieniem kanału). Przy takiej kolejności każdy użytkownik najpierw dekoduje s_K, a następnie odejmuje składową t_s od otrzymanego sygnału. W rezultacie typowy użytkownik i po pierwszej iteracji SIC otrzymuje

$$\bar{r}_i = r_i - \sqrt{\alpha_K P_d} g_i s_K = g_i \sum_{k=1}^{K-1} \sqrt{\alpha_k P_d} s_k + n_i \quad (138)$$

zakładając bezbłędne wykrywanie i doskonałą znajomość kanału. W drugiej iteracji każdy użytkownik dekoduje s_{K-1} przy użyciu pozostałego sygnału $\bar{r}_i$, bez zakłóceń ze strony najsłabszego użytkownika. Anulowanie powtarza się, aż każdy użytkownik otrzyma swój własny sygnał. W szczególności najsłabszy

użytkownik dekoduje swój własny sygnał bezpośrednio bez kolejnych anulowań zakłóceń, ponieważ przydzielono mu największą moc. Dlatego SNRdla użytkownika K można zapisać jako

$$\gamma_K = \frac{|g_K|^2 \alpha_K P_d}{|g_K|^2 \sum_{k=1}^{K-1} \alpha_k P_d + N_0 B_w}, \quad (139)$$

gdzie B_w oznacza szerokość pasma sygnału. Ogólnie rzecz biorąc, SNR dla użytkownika i wynosi

$$\gamma_i = \frac{|g_i|^2 \alpha_i P_d}{|g_i|^2 \sum_{k=1}^{i-1} \alpha_k P_d + N_0 B_w}, \quad (140)$$

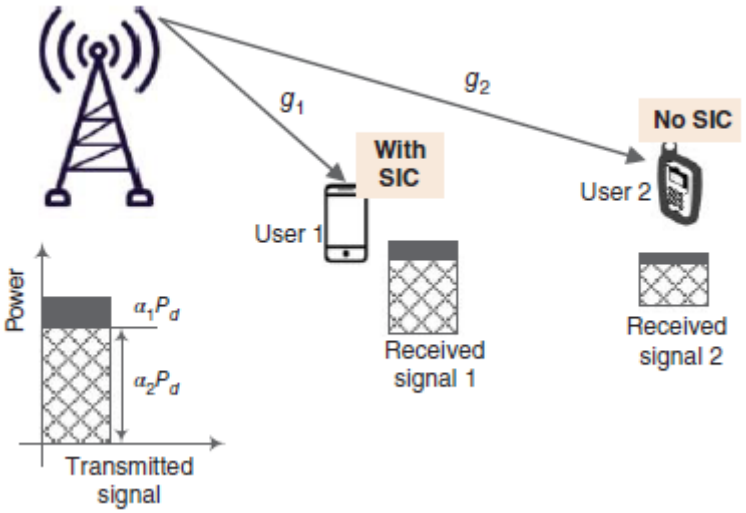
co skutkuje osiągalną stawką

$$R_i = B_w \log\left(1 + \frac{|g_i|^2 \alpha_i P_d}{|g_i|^2 \sum_{k=1}^{i-1} \alpha_k P_d + N_0 B_w}\right) \quad (141)$$

Suma szybkości transmisji DL NOMA jest obliczana na podstawie

$$R = \sum_{i=1}^{K} R_i. \quad (142)$$

W przypadku dwóch użytkowników, jak pokazano na rysunku,



dalekiemu użytkownikowi na skraju komórki przydzielana jest większa moc, a bliskiemu użytkownikowi w centrum komórki przydzielana jest mniejsza moc. Niezależnie od różnicy g_1 i g_2, ten stosunek mocy będzie zachowany w odebranym sygnale dowolnego użytkownika. Bliski użytkownik najpierw wykrywa symbol dalekiego użytkownika s_2, a następnie odejmuje jego zregenerowany składnik od odebranego sygnału, co skutkuje dostępną szybkością

$$R_1 = \log\left(1 + \frac{|g_1|^2 \alpha_1 P_d}{N_0}\right) \quad (143)$$

zakładając, że szerokość pasma sygnału jest znormalizowana do B_w = 1 Hz. Daleki użytkownik wykrywa swój sygnał bezpośrednio, traktując sygnał bliskiego użytkownika jako kolorowy szum, uzyskując

$$R_2 = \log\left(1 + \frac{|g_2|^2\alpha_2 P_d}{|g_2|^2\alpha_1 P_d + N_0}\right) \quad (144)$$

W rezultacie transmisja DL między dwoma użytkownikami przy użyciu NOMA uzyskuje sumaryczną szybkość

$$R = R_1 + R_2 = \log\left(1 + \frac{|g_1|^2\alpha_1 P_d}{N_0}\right) + \log\left(1 + \frac{|g_2|^2\alpha_2 P_d}{|g_2|^2\alpha_1 P_d + N_0}\right) \quad (145)$$

Używając schematu OMA jako porównania, bliskiemu użytkownikowi przypisuje się pasmo η Hz ($0 < \eta < 1$), pozostawiając resztę pasma $(1 - \eta)$ Hz dalekiemu użytkownikowi. Ich osiągalne szybkości można obliczyć za pomocą

$$R_1 = \eta \log\left(1 + \frac{|g_1|^2\alpha_1 P_d}{\eta N_0}\right)$$
$$R_2 = (1-\eta)\log\left(1 + \frac{|g_2|^2\alpha_2 P_d}{(1-\eta)N_0}\right) \quad (146)$$

Załóżmy, że geometrie bliskiego użytkownika i dalekiego użytkownika wynoszą odpowiednio $|g_1|^2P_d/N_0$ = 20 dB i $|g_2|^2P_d/N_0$ = 0 dB. Gdy równe pasmo (tj. η = 0,5) i równa moc (tj. $\alpha_1 = \alpha_2$ = 0,5) są przydzielane każdemu użytkownikowi z kryteriami proporcjonalnej sprawiedliwości, stawki OMA na użytkownika są obliczane zgodnie z równaniem (146) jako

$$R_1 = 0.5\log_2(1 + 100) = 3.3291 \text{ bps/Hz}$$
$$R_2 = 0.5\log_2(1 + 1) = 0.5 \text{ bps/Hz}. \quad (147)$$

Z drugiej strony, gdy przydział mocy w NOMA jest przeprowadzany jako α_1 = 0,2 i α_2 = 0,8, stawki na użytkownika są obliczane zgodnie z równaniami (143) i (144) jako

$$R_1 = \log_2(1 + 20) = 4.3923 \text{ bps/Hz}$$
$$R_2 = \log_2(1 + 0.8/1.2) = 0.7370 \text{ bps/Hz}, \quad (148)$$

co odpowiada zyskowi wydajności widmowej wynoszącemu około 32% i 47% w odniesieniu do schematu OMA. Taki zysk jest zbierany poprzez pełne wykorzystanie różnicy wzmocnień kanałów między użytkownikami. Większa różnica zwykle odpowiada wyższemu zyskowi wydajności widmowej NOMA w porównaniu z OMA i odwrotnie. Jeśli dwóch użytkowników ma identyczne warunki kanału, tj. $|g_1| = |g_2|$, sumaryczną szybkość NOMA w równaniu (145) można zapisać jako

$$R = \log\left(1 + \frac{|g_1|^2\alpha_1 P_d}{N_0}\right) + \log\left(1 + \frac{|g_2|^2\alpha_2 P_d}{|g_2|^2\alpha_1 P_d + N_0}\right)$$
$$= \log\left(1 + \frac{|g_1|^2 P_d}{N_0}\right) = \log\left(1 + \frac{|g_2|^2 P_d}{N_0}\right), \quad (149)$$

co jest dokładnie identyczne z sumaryczną szybkością OMA. Oznacza to, że zysk wydajności NOMA znika, jeśli użytkownicy mają takie same lub podobne warunki kanału.

Uplink Non-Orthogonal Multiple Access

Transmisja UL NOMA nieznacznie różni się od jej odpowiednika DL, gdzie K rozproszonych przestrzennie użytkowników wyposażonych w pojedynczą antenę jednocześnie przesyła swoje symbole nośne informacji w kierunku stacji bazowej z pojedynczą anteną przez tę samą jednostkę zasobów. Odebrany sygnał w stacji bazowej można wyrazić za pomocą

$$r = \sum_{k=1}^{K} g_k \sqrt{P_k} s_k + n, \quad (150)$$

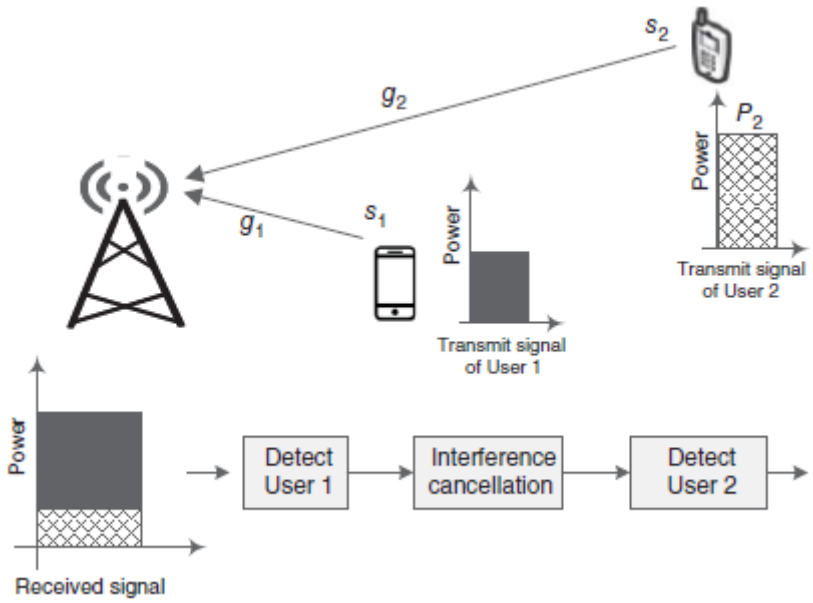
gdzie s_k jest symbolem nośnym informacji użytkownika k, k = 1, 2,…, K, spełniającym $\mathbb{E}[|s_k|^2]= 1$, a P_k oznacza ograniczenie mocy użytkownika k, g_k oznacza zespolone wzmocnienie kanału od użytkownika k do stacji bazowej, a n jest addytywnym białym szumem gaussowskim o zerowej średniej i gęstości widmowej mocy N_0 (W/Hz). Podobnie możemy również założyć, że użytkownik 1 ma największe wzmocnienie kanału, a użytkownik K jest najsłabsze, tj.

$$|g_1| \geqslant |g_2| \geqslant \cdots \geqslant |g_K|, \quad (151)$$

Jako rodzaj techniki domeny mocy, przydział mocy odgrywa kluczową rolę w DL NOMA, która ma co najmniej dwie główne funkcje: sztuczne tworzenie wystarczającej różnicy mocy w odebranym sygnale, aby ułatwić kolejne usuwanie zakłóceń i gwarantowanie rozsądnej mocy odebranej dla użytkowników na krawędzi komórki. W UL użytkownicy mogą ponownie optymalizować swoje moce nadawcze zgodnie ze swoimi lokalizacjami, tak jak w DL. Jednak użytkownicy mogą być dobrze rozproszeni w zasięgu komórki, a poziomy mocy odebranej od różnych użytkowników są już dobrze rozdzielone. Stacja bazowa może najpierw zdekodować symbol użytkownika o największej odebranej mocy, traktując sygnały wszystkich innych użytkowników jako kolorowy szum. Następnie anuluje odpowiednie zakłócenia z odebranego sygnału i kontynuuje dekodowanie symbolu innego użytkownika o drugiej co do wielkości odebranej mocy. Proces SIC powtarza się na stacji bazowej, aż do wykrycia wszystkich symboli. Przestrzegając kolejności dekodowania 1 → 2 → · · · → K, można obliczyć możliwą do osiągnięcia szybkość typowego użytkownika k, korzystając z następującego wzoru:

$$R_k = B_w \log\left(1 + \frac{|g_k|^2 P_k}{\sum_{i=k+1}^{K} |g_i|^2 P_i + N_0 B_w}\right) \quad (152)$$

W przypadku dwóch użytkowników, jak pokazano na rysunku 10.16,



użytkownik bliski w centrum komórki nadaje s_1 z mocą P_1 w kierunku stacji bazowej, podczas gdy użytkownik daleki na skraju komórki jednocześnie nadaje s_2 z mocą P_2. Ich moc może być identyczna lub użytkownik daleki ma większą moc, aby zagwarantować rozsądną moc odbieranego sygnału. W praktyce różnica g_1 i g_2 powoduje wystarczającą różnicę mocy między składowymi odbieranego sygnału tych dwóch użytkowników, niezależnie od ich przerwy w mocy transmisji. Stacja bazowa może najpierw wykryć użytkownika bliskiego bezpośrednio bez SIC, co skutkuje

$$R_1 = \log\left(1 + \frac{|g_1|^2 P_1}{|g_2|^2 P_2 + N_0}\right) \quad (153)$$

zakładając, że szerokość pasma sygnału jest znormalizowana do B_w = 1 Hz. Następnie stacja bazowa anuluje zakłócenia pierwszego użytkownika i wykrywa sygnał drugiego użytkownika, uzyskując

$$R_2 = \log\left(1 + \frac{|g_2|^2 P_2}{N_0}\right) \quad (154)$$

Suma szybkości transmisji UL NOMA dla dwóch użytkowników jest obliczana na podstawie

$$\begin{aligned} R &= R_1 + R_2 \\ &= \log\left(1 + \frac{|g_1|^2 P_1}{|g_2|^2 P_2 + N_0}\right) + \log\left(1 + \frac{|g_2|^2 P_2}{N_0}\right) \\ &= \log\left(1 + \frac{|g_1|^2 P_1 + |g_2|^2 P_2}{N_0}\right). \end{aligned} \quad (155)$$

Jeżeli stacja bazowa najpierw bezpośrednio wykryje dalekiego użytkownika, a następnie wykryje bliskiego użytkownika bez zakłóceń między użytkownikami, ich osiągalne szybkości stają się

$$\begin{aligned} R_1 &= \log\left(1 + \frac{|g_1|^2 P_1}{N_0}\right) \\ R_2 &= \log\left(1 + \frac{|g_2|^2 P_2}{|g_1|^2 P_1 + N_0}\right) \end{aligned} \quad (156)$$

Co ciekawe, jeśli nie weźmiemy pod uwagę wyższego prawdopodobieństwa błędu wynikającego z niskiego stosunku sygnału do szumu (SNR) przy pierwszym wykryciu dalekiego użytkownika, sumaryczna szybkość transmisji UL NOMA między dwoma użytkownikami jest taka sama niezależnie od kolejności SIC, ponieważ

$$\begin{aligned} R &= R_1 + R_2 \\ &= \log\left(1 + \frac{|g_1|^2 P_1}{N_0}\right) + \log\left(1 + \frac{|g_2|^2 P_2}{|g_1|^2 P_1 + N_0}\right) \\ &= \log\left(1 + \frac{|g_1|^2 P_1 + |g_2|^2 P_2}{N_0}\right) \end{aligned} \tag{157}$$

dokładnie równa się równaniu (155). W konsekwencji transmisja UL dla dwóch użytkowników z NOMA tworzy następujący obszar pojemności

$$\mathfrak{C} = \left\{ (R_1, R_2) \in \mathbb{R}_+^2 \;\middle|\; \begin{array}{l} R_1 \leqslant \log\left(1 + \frac{|g_1|^2 P_1}{N_0}\right) \\ R_2 \leqslant \log\left(1 + \frac{|g_2|^2 P_2}{N_0}\right) \\ R_1 + R_2 \leqslant \log\left(1 + \frac{|g_1|^2 P_1 + |g_2|^2 P_2}{N_0}\right) \end{array} \right\} \tag{158}$$

W schemacie OMA pierwszy użytkownik zajmuje η całkowitych zasobów czasowo-częstotliwościowych, pozostawiając resztę $(1 - \eta)$ zasobów czasowo-częstotliwościowych drugiemu użytkownikowi. Ich osiągalne stawki można obliczyć za pomocą

$$\begin{aligned} R_1 &= \eta \log\left(1 + \frac{|g_1|^2 P_1}{\eta N_0}\right) \\ R_2 &= (1 - \eta) \log\left(1 + \frac{|g_2|^2 P_2}{(1 - \eta) N_0}\right) \end{aligned} \tag{159}$$

Załóżmy, że geometrie bliskiego użytkownika i dalekiego użytkownika to odpowiednio $|g_1|^2 P_1/N_0 = 20$ dB i $|g_2|^2 P_2/N_0 = 0$ dB. Gdy równe pasmo (tj. $\eta = 0{,}5$) jest przydzielone każdemu użytkownikowi z kryteriami proporcjonalnej uczciwości, stawki OMA na użytkownika są obliczane zgodnie z równaniem (159) jako

$$\begin{aligned} R_1 &= 0.5 \log_2(1 + 200) = 3.8255 \text{ bps/Hz} \\ R_2 &= 0.5 \log_2(1 + 2) = 0.7925 \text{ bps/Hz}. \end{aligned} \tag{160}$$

Natomiast stawki transmisji UL NOMA na użytkownika obliczane są według równań (153) i (154) (najpierw dekodowanie najbliższego użytkownika) jako

$$\begin{aligned} R_1 &= \log_2(1 + 50) = 5.6724 \text{ bps/Hz} \\ R_2 &= \log_2(1 + 1) = 1 \text{ bps/Hz}, \end{aligned} \tag{161}$$

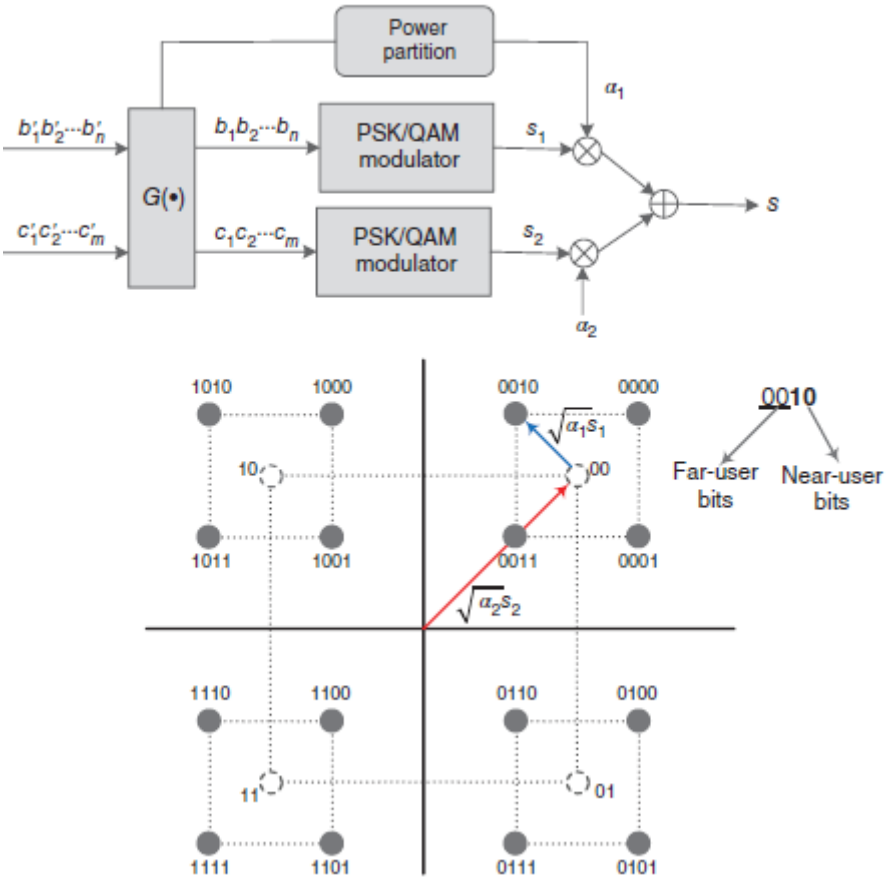
co odpowiada zyskowi wydajności widmowej wynoszącemu około 48% i 26% w odniesieniu do schematu OMA. Odwracając kolejność SIC, stawki transmisji UL NOMA na użytkownika obliczane są zgodnie z równaniem (156) (najpierw dekodując dalekiego użytkownika) jako

$$R_1 = \log_2(1 + 100) = 6.6582 \text{ bps/Hz}$$
$$R_2 = \log_2(1 + 1/101) = 0.0142 \text{ bps/Hz.} \quad (162)$$

Wyniki te odpowiadają zyskowi wydajności widmowej wynoszącemu około 74% dla bliskiego użytkownika, ale stracie wynoszącej −98% dla dalekiego użytkownika, co oznacza znaczenie uporządkowania SIC. Można zauważyć, że różne uporządkowania SIR dają tę samą sumaryczną szybkość R = 6,6724 bps/Hz, co potwierdza równoważność równań (155) i (157).

Kodowanie superpozycji wielu użytkowników

Dzięki rozwojowi implementacji odbiornika i możliwości sprzętowych, usuwanie zakłóceń staje się bardziej przystępne cenowo w terminalach mobilnych. Sprawia, że transmisja nieortogonalna jest bardziej wykonalna. W DL, NOMA jest obiecującą techniką zwiększania przepustowości systemu i poprawy doświadczeń użytkownika, np. transmisja superpozycji wielu użytkowników (MUST) określona w 3GPP dla usług szerokopasmowych DL. Zasadniczo wielu użytkowników można multipleksować w każdej ortogonalnej jednostce zasobów. Jednak biorąc pod uwagę malejące zyski przy superpozycji większej liczby użytkowników i narzut sygnalizacyjny, superpozycja dwóch użytkowników jest zazwyczaj skoncentrowana. Najprostszą metodą jest superpozycja liniowa, niezależnie mapująca zakodowane bity dwóch lub więcej współzaplanowanych użytkowników na symbole konstelacji składowych, które są superponowane z adaptacyjnym współczynnikiem mocy. Rysunek pokazuje kodowanie bezpośredniej superpozycji nad dwoma użytkownikami, gdzie n i m zakodowanych bitów bliskich i dalekich użytkowników, oznaczonych odpowiednio jako b_1, b_2,…, b_n i c_1, c_2,…, c_n, jest przesyłanych jednocześnie.



Bity są oddzielnie modulowane za pomocą starszych modulatorów Phase-Shift Keying (PSK) lub Quadrature Amplitude Modulation (QAM), z których każdy ma 2^n i 2^m punktów konstelacji. Stosowany jest elastyczny podział mocy między bliskim i dalekim użytkownikiem, oznaczonym jako $\alpha_1 P_d$ i $\alpha_2 P_d$. Stosunek mocy α_1/α_2 jest wybierany w celu zmaksymalizowania szybkości sumy przy określonym kryterium uczciwości, przy jednoczesnym uwzględnieniu wzmocnień kanału itd. Po przeskalowaniu za pomocą odpowiednich współczynników mocy, dwa symbole modulacji, s_1 i s_2, są nakładane na siebie, co skutkuje

$$s = \sqrt{\alpha_1 P_d}s_1 + \sqrt{\alpha_2 P_d}s_2. \quad (163)$$

Symbol złożony s ma konstelację z 2^{n+m} punktów. 16-punktowa konstelacja na rysunku jest przykładem, gdy QPSK jest stosowane zarówno dla użytkowników bliskich, jak i dalekich, gdzie $\alpha_1 < 0{,}5$, ponieważ użytkownik bliski jest przypisany do małej części mocy, a amplituda jego symbolu jest zatem mała. Symbol użytkownika dalekiego ma większą moc, co decyduje, w którym kwadrancie będzie leżał symbol złożony. Innymi słowy, punkt konstelacji użytkownika bliskiego zależy od punktu konstelacji użytkownika dalekiego. Jako przykład pokazany na rysunku, zakodowane bity dalekiego użytkownika 00 najpierw decydują o pierwszym kwadrancie, a następnie jeden z czterech punktów w pierwszym kwadrancie jest wybierany zgodnie z bitami kodów bliskiego użytkownika 10. Chociaż konstelacja każdego użytkownika podąża za mapowaniem Graya (tj. dwa sąsiednie punkty konstelacji różnią się tylko jedną cyfrą binarną), konstelacja bezpośredniej superpozycji nie jest rodzajem mapowania Graya, co prowadzi do utraty przepustowości. Konstelacja niemapowana według Graya w dużym stopniu opiera się na zaawansowanym odbiorniku. W tym przypadku anulowanie zakłóceń na poziomie

symboli nie może wystarczająco rozwiązać rozmytej konstelacji z powodu zakłóceń między użytkownikami. Bardziej skomplikowane odbiorniki, np. anulowanie zakłóceń na poziomie słów kodowych, są potrzebne do osiągnięcia akceptowalnej wydajności. Mapowanie Graya zapewnia solidną wydajność nawet wtedy, gdy anulowanie zakłóceń jest wykonywane na poziomie symboli bez dekodowania kanału, co jest zatem znacznie prostsze niż anulowanie zakłóceń na poziomie słów kodowych. Jak pokazano na rysunku , wprowadzanie zakodowanych bitów do konwertera bitów oznaczonego jako G(·) może zapewnić właściwość Graya w złożonej konstelacji. Alternatywnie, schemat oparty na partycjonowaniu bitów, a nie partycjonowaniu mocy, może zostać zastosowany do wdrożenia mapowania Graya, gdzie konstelacja jest legacy QAM mapper z równomiernie rozmieszczoną prostokątną siatką, a zakodowane bity dwóch lub więcej użytkowników są bezpośrednio nałożone na symbole złożonej konstelacji [Yifei, 2016]. Element badania 3GPP, „Badanie nad transmisją superpozycji DL Multiuser", został przeprowadzony w 3GPP w celu oceny wydajności systemu potencjalnych ulepszeń LTE umożliwiających DL MUST. Cele obejmują zdefiniowanie scenariuszy wdrożenia docelowego i metodologii oceny dla MUST, identyfikację potencjalnych schematów MUST i odpowiadających im ulepszeń LTE oraz ocenę wykonalności i wydajności na poziomie systemu możliwych schematów MUST. W rezultacie 3GPP zaleciło w 3GPP TR36.859 [2015] trzy różne kategorie MUST:

● Kategoria MUST 1: Transmisja superpozycji z adaptacyjnym współczynnikiem mocy w konstelacjach składowych i konstelacji złożonej bez mapowania Graya, gdzie zakodowane bity dwóch lub więcej współplanowanych użytkowników są niezależnie mapowane na symbole konstelacji składowych, ale konstelacja złożona nie ma mapowania Graya.

● Kategoria MUST 2: Transmisja superpozycji z adaptacyjnym współczynnikiem mocy w konstelacjach składowych i konstelacji złożonej z mapowaniem Graya, gdzie zakodowane bity dwóch lub więcej współplanowanych użytkowników są wspólnie mapowane na konstelacje składowe, a następnie konstelacja złożona ma mapowanie Graya.

● Kategoria MUST 3: Transmisja superpozycji z przypisaniem bitów etykiet w konstelacji złożonej z mapowaniem Graya, gdzie zakodowane bity dwóch lub więcej współplanowanych użytkowników są bezpośrednio mapowane na symbole konstelacji złożonej.

W związku z tym 3GPP zasugerowało kilka schematów odbiorników kandydackich, aby umożliwić różne schematy MUST zarówno dla użytkowników dalekich, jak i bliskich. Obiecujące schematy odbiorników dla użytkowników dalekich obejmują:

● Liniowy MMSE z odrzucaniem zakłóceń łączący

● Wykrywanie o maksymalnym prawdopodobieństwie

● Wykrywanie o zmniejszonej złożoności o maksymalnym prawdopodobieństwie

● Anulowanie zakłóceń na poziomie symboli

Schematy odbiorników dla użytkowników bliskich mogą być:

● Wykrywanie o maksymalnym prawdopodobieństwie

● Wykrywanie o zmniejszonej złożoności o maksymalnym prawdopodobieństwie

● Anulowanie zakłóceń na poziomie symboli

● Liniowe anulowanie kolejnych zakłóceń na poziomie słowa kodowego

● Anulowanie kolejnych zakłóceń na poziomie słowa kodowego

Transmisja bez przyznawania uprawnień w łączu w górę

Większość transmisji UL w tradycyjnych systemach mobilnych planuje użytkowników ortogonalnie, z dedykowaną jednostką zasobów czasowo-częstotliwościowych na użytkownika. Takie planowanie użytkowników prowadzi do dużego narzutu sygnalizacyjnego między siecią a terminalami. Z drugiej strony ładunek danych jest zazwyczaj nierównomierny i mały w scenariuszach wdrażania nowoczesnego Internetu rzeczy, ale liczba połączeń jest ogromna. Wysokie zużycie energii przez urządzenia ze względu na znaczny narzut na użytkownika i ścisły mechanizm kontroli sprawiają, że projekt taniego terminala o niskim poborze mocy staje się trudniejszy. NOMA umożliwia jednoczesną obsługę większej liczby użytkowników i ułatwia transmisję bez przyznawania uprawnień, dzięki czemu system nie jest ściśle ograniczony liczbą zasobów ortogonalnych i ich granularnością harmonogramowania. Aby złagodzić efekt kolizji zasobów w transmisji nieortogonalnej, można zastosować rozpraszanie. Historycznie, transmisja nieortogonalna była używana w transmisji UL IS-95, CDMA2000 i WCDMA (Wideband Code Division MultipleAccess), aby wymienić kilka, które głównie obsługują użytkowników głosowych komutowanych obwodami z ciągłym, ale małym strumieniowaniem danych. Systemy te opierają się na bezpośrednim sekwencyjnym widmie rozproszonym, umożliwiając wielu użytkownikom współdzielenie wspólnego zasobu czasowo-częstotliwościowego za pomocą zestawu kodów rozpraszających. Podobnie, koncepcja rozprzestrzeniania może być używana w UL do masowej łączności z bardziej zaawansowanymi technikami. Podnosi to nową filozofię projektowania transmisji nieortogonalnej określanej jako NOMA domeny kodu, w porównaniu z konwencjonalnym projektem zwanym NOMA domeny zasilania. Transmisja UL bez przyznawania grantów składa się z następujących krytycznych składników

● Rozprzestrzenianie kodu: Jak pokazuje technologia 3G oparta na bezpośrednim sekwencyjnym widmie rozproszenia, rozprzestrzenianie może poprawić odporność systemu na zakłócenia współkanałowe i międzyużytkownikowe. Graf czynnikowy jest skutecznym narzędziem optymalizacji projektu. Na grafie czynnikowym wiele węzłów zmiennych jest połączonych z wieloma węzłami czynnikowymi. Połączenia między węzłami zmiennymi i węzłami czynnikowymi definiują kluczową właściwość schematów dostępu nieortogonalnego. Ponadto mapa połączeń dostarcza wskazówek dotyczących implementacji odbiornika, np. rozprzestrzeniania opartego na sygnaturach o niskiej gęstości, co może obniżyć złożoność wykrywania poprzez uniknięcie obliczeń metrycznych pełnych połączeń między węzłami zmiennymi i węzłami czynnikowymi. Ponadto złożone sekwencje niebinarne mają niższą korelację krzyżową między różnymi sekwencjami w porównaniu z sekwencjami binarnymi, nawet gdy są bardzo krótkie. Te cechy mogą ułatwić zakwaterowanie znacznie bardziej aktywnych użytkowników w zasobach współdzielonych, gdy użytkownicy ci losowo wybierają sekwencje rozprzestrzeniania.

● Tryb działania: W transmisji bez przyznawania uprawnień adaptacja łącza jest wykonywana w stylu długoterminowym. Długoterminowo oznacza to, że wybór schematu modulacji i kodowania (MSC) zależy tylko od zaniku sygnału na dużą skalę i sterowania mocą w pętli otwartej. Schemat kodowania modulacji (MCS) nie jest często dostosowywany, ponieważ wahania zaniku sygnału na dużą skalę zmieniają się stosunkowo wolno, a nacisk harmonogramowania bez przyznawania nie jest położony na maksymalizację pojemności systemu.

● Projekt odbiornika: Dwa rodzaje odbiorników zyskały dużą uwagę:

– Eliminacja zakłóceń na poziomie bitów

– Algorytm przekazywania wiadomości (MPA)

Anulowanie zakłóceń na poziomie bitów jest podobne do anulowania zakłóceń na poziomie słów kodowych. W przypadku UL anulowanie zakłóceń na poziomie bitów staje się bardziej przystępne cenowo niż w przypadku DL, ponieważ stacja bazowa musi zdekodować bity wszystkich aktywnych urządzeń. MPA to suboptymalny algorytm wykrywania grafu czynnikowego o maksymalnym prawdopodobieństwie. Proces wykrywania jest iteracyjny, podobny do procesu dekodera LDPC (Low Density ParityCheck), w którym informacje dotyczące przekonań lub informacje zewnętrzne przepływają w obie strony między węzłami zmiennymi i węzłami czynnikowymi. Zaprojektowano różne schematy dla UL NOMA, aby obsługiwać łączność masową i umożliwiać transmisję bez przyznawania uprawnień z niskim opóźnieniem. Na przykład, w sumie 15 propozycji zostało złożonych do 3GPP podczas projektowania transmisji New Radio UL , w tym implementacja zarówno domeny kodu, jak i domeny mocy, tj. Sparse Code Multiple Access (SCMA), Multi-User Shared Access (MUSA), Low Code Rate Spreading, Frequency-Domain Spreading, Non-orthogonal Coded Multiple Access (NCMA), NOMA, Pattern Division Multiple Access (PDMA), Resource Spread Multiple Access (RSMA), Interleave-Grid Multiple Access (IGMA), Low Density Spreading with Signature Vector Extension (LDS-SVE), Low code rate and Signature-based Shared Access (LSSA), Non-Orthogonal Coded Access (NOCA), Interleave-Division Multiple Access (IDMA), Repetition-Division Multiple Access (RDMA) i Group Orthogonal Coded Access (GOCA). Schematy te mają wspólną podstawę i wiele podobieństw niezależnie od ich konkretnych właściwości. Aby zapewnić wgląd, w kolejnej części przeanalizujemy zasady i aspekty projektowe kilku typowych schematów NOMA domeny kodowej.

NOMA domeny kodowej

Oprócz różnicowania sygnałów różnych użytkowników w domenie mocy, praktyczną implementację NOMA można również zrealizować w domenie kodowej. Najbardziej reprezentatywne schematy NOMA domeny kodowej obejmują CDMA lub OFDM oparte na Low-Density Signature (LDS) oraz SCMA. Ta część pokrótce przedstawi ich podstawowe zasady i główne cechy, aby dać czytelnikom wgląd w te oparte na rozprzestrzenianiu nieortogonalne schematy transmisji.

CDMA/OFDM z Low-Density Signature

W konwencjonalnym CDMA nie można osiągnąć ortogonalnej kanałowości, gdy liczba aktywnych użytkowników jest większa niż zysk przetwarzania, a mianowicie stan przeciążenia. Używając gęstej struktury gęstości, każdy chip otrzymanego sygnału zawiera wkład wszystkich współzaplanowanych użytkowników w systemie. Innymi słowy, każdy użytkownik cierpi z powodu zakłóceń wielodostępowych od wszystkich innych użytkowników na każdym chipie. Jeśli macierz korelacji krzyżowej podpisu jest zgodna z pewnym formatem, złożoność optymalnych algorytmów MUD może zostać obniżona, ale jej złożoność jest nadal zbyt wysoka, aby stała się przystępna cenowo. Zainspirowani sukcesem struktury o niskiej gęstości w kodach LDPC, Hoshyar i inNI zaproponowali nową transmisję CDMA opartą na LDS. Wraz z iteracyjnym algorytmem przekazywania wiadomości na poziomie układu scalonego lub wykrywaniem opartym na MPA, jej osiągalna wydajność zbliża się do wydajności systemu pojedynczego użytkownika ze współczynnikiem przeciążenia wynoszącym 200% przy przystępnej złożoności obliczeniowej, w porównaniu z konwencjonalną strukturą wykorzystującą optymalny MUD. Rozważmy system UL CDMA, w którym K synchronicznych użytkowników jednocześnie przesyła swoje symbole x_k, k = 1, 2,…, K w kierunku stacji bazowej za pomocą zestawu sekwencji rozpraszających o długości N. Modulowany symbol x_k jest tworzony przez mapowanie sekwencji niezależnych bitów informacji na alfabet konstelacji. Następnie zmodulowany symbol jest multipleksowany z sekwencją rozprzestrzeniania $s_k = [s_{1,k}, s_{2,k},..., s_{N,k}]^T$, przypisaną unikalnie każdemu użytkownikowi. W konwencjonalnej strukturze CDMA każdy składnik sekwencji rozprzestrzeniania zwykle przyjmuje wartości różne od zera, co jest optymalizowane do pewnych ograniczeń, np. wysokiej

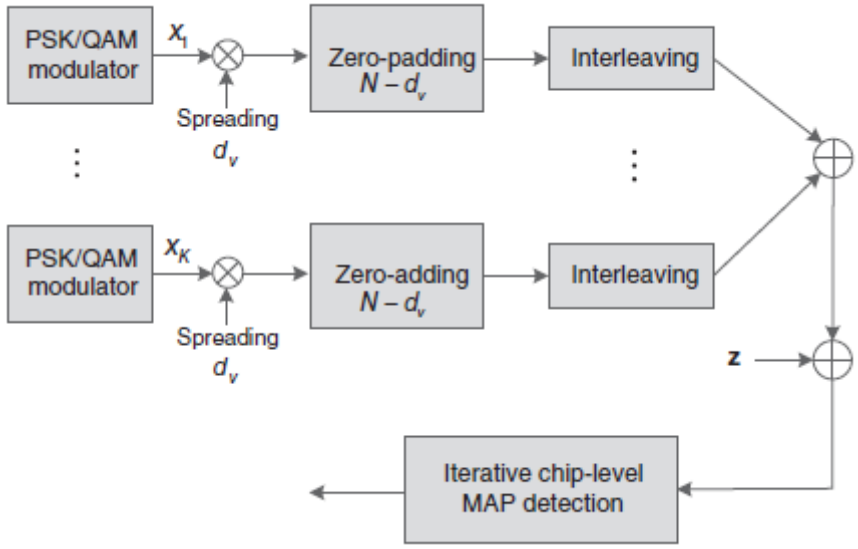
autokorelacji i niskiej relacji krzyżowej. Następnie sygnał otrzymany na chipie n, n = 1, 2,…,N można wyrazić za pomocą

$$y_n = \sum_{k=1}^{K} g_k s_{n,k} x_k + z_n, \quad (164)$$

gdzie g_k oznacza wzmocnienie kanału między użytkownikiem k a stacją bazową, $s_{n,k}$ jest n-tym składnikiem ciągu rozprzestrzeniania s_k, a z_n oznacza addytywny szum gaussowski. Układając N kolejnych chipów symbol po symbolu, otrzymany wektor sygnału $y = [y_1, y_2,\ldots, y_N]^T$ jest superpozycją sygnałów przesyłanych od wszystkich użytkowników, którą można sformułować jako

$$\mathbf{y} = \sum_{k=1}^{K} g_k \mathbf{s}_k x_k + \mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{z}, \quad (165)$$

gdzie $x = [x_1, x_2,\ldots, x_K]^T$, $z = [z_1, z_2,\ldots, z_N]^T$, a efektywny podpis odbioru $H = [g_1s_1, g_2s_2,\ldots, g_Ks_K]$. Zamiast optymalizować sygnatury N-chipów, struktura LDS celowo organizuje każdego użytkownika tak, aby rozłożył swój zmodulowany symbol na niewielką liczbę chipów d_v, po czym następuje proces wypełniania zerami tak, aby wzmocnienie przetwarzania nadal wynosiło N, jak pokazano na rysunku.



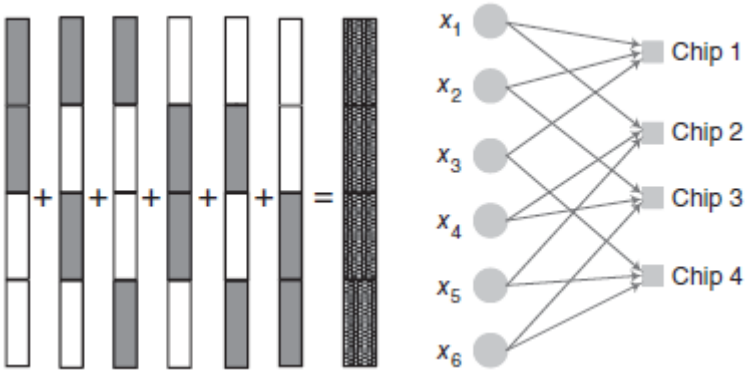
Ponadto niech d_c będzie maksymalną liczbą użytkowników, którym wolno zakłócać pojedynczy chip. Sekwencje rozłożone i wypełnione są następnie przeplatane unikalnie dla każdego użytkownika tak, że wynikowa macierz sygnatur staje się rzadka. Tę strukturę o niskiej gęstości można wyrazić za pomocą macierzy wskaźników, gdzie wpis 1 w n-tym wierszu i k-tej kolumnie oznacza, że użytkownik k rozprzestrzenia swój sygnał na chipie n, gdzie 0 oznacza, że użytkownik k wyłącza się na tym chipie. Zbiór pozycji jedynek w n-tym wierszu macierzy wskaźników oznacza zbiór użytkowników, którzy wnoszą swoje dane w n-tym chipie, podczas gdy jej k-ta kolumna reprezentuje zbiór chipów, na których użytkownik k rozprzestrzenia swoje dane. Niech ξ_n i ζ_k będą zbiorem pozycji jedynek w n-tym wierszu i k-tej kolumnie, odpowiednio, równanie (164) można zapisać jako

$$y_n = \sum_{k \in \xi_n^g} g_k s_{n,k} x_k + z_n. \quad (166)$$

W rezultacie liczba nałożonych sygnałów na każdym chipie będzie mniejsza niż liczba aktywnych użytkowników, co zmniejszy zakłócenia spowodowane przez wielu użytkowników. Na przykład macierz wskaźników

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}, \quad (167)$$

oznacza ustawienie, w którym K = 6 użytkowników nakłada swoje symbole x_k, k = 1,…,6 na N = 4 chipy. Ponieważ liczba użytkowników jest większa niż liczba ortogonalnych jednostek zasobów, osiąga się NOMA. Każdy użytkownik rozprzestrzenia swój symbol, używając unikalnej sekwencji rozprzestrzeniania 4-chipów składającej się tylko z dwóch niezerowych komponentów, tj. $d_v = 2$. Każdy chip obsługuje tylko trzech użytkowników, tj. $d_c = 3$, a nie wszystkich sześciu użytkowników, dzięki czemu zmniejszono interferencję wielokrotnego dostępu. Tymczasem operacja rozprzestrzeniania może być wyrażona za pomocą grafu czynnikowego, który zawiera liczbę zmiennych węzłów i węzłów czynnikowych, jak pokazano na rysunku.



Zmienne węzły zazwyczaj reprezentują modulowane symbole lub zakodowane bity, a węzły czynnikowe oznaczają ortogonalne jednostki zasobów czasowo-częstotliwościowych. Połączenie między węzłami zmiennymi i czynnikowymi może rozróżniać nieortogonalne schematy dostępu. Węzeł zmienny jest w stanie połączyć się z wieloma węzłami czynnikowymi, co jest w rzeczywistości procesem rozprzestrzeniania. Węzeł czynnikowy może połączyć się z wieloma węzłami zmiennymi, co oznacza nieortogonalną alokację zasobów. Zastosowanie rozprzestrzeniania o niskiej gęstości w OFDM jest proste, jeśli zastąpimy każdy chip podnośną OFDM, tworząc nową technikę zwaną Low-Density Structure-OFDM lub LDS-OFDM. Należy zauważyć, że ograniczeniem projektowym LDS-CDMA i LDS-OFDM jest to, że liczba użytkowników jest większa niż liczba chipów (lub podnośnych), mianowicie K > N, aby osiągnąć NOMA. Tymczasem liczba użytkowników odpowiadająca każdemu chipowi lub podnośnej powinna być mniejsza niż liczba chipów lub podnośnych, np. $d_c < N$ gwarantuje, że interferencja wielodostępowa zostanie obniżona w porównaniu ze standardową strukturą o dużej gęstości.

Sparse Code Multiple Access

SCMA to ulepszona wersja podstawowego LDS-CDMA. Podstawowa zasada LDS-CDMA i SCMA jest taka sama: wykorzystanie niskiej gęstości (rzadkiej sekwencji składowych niezerowych) w celu zmniejszenia złożoności wykrywania MPA w odbiorniku. Kluczowym pomysłem SCMA jest połączenie mapowania konstelacji i rozprzestrzeniania w celu mapowania zakodowanych bitów bezpośrednio na słowo kodowe. Cały proces można interpretować jako procedurę kodowania z domeny binarnej do złożonej domeny wielowymiarowej. SCMA został zaproponowany w Nikopour i Baligh z następującymi właściwościami:

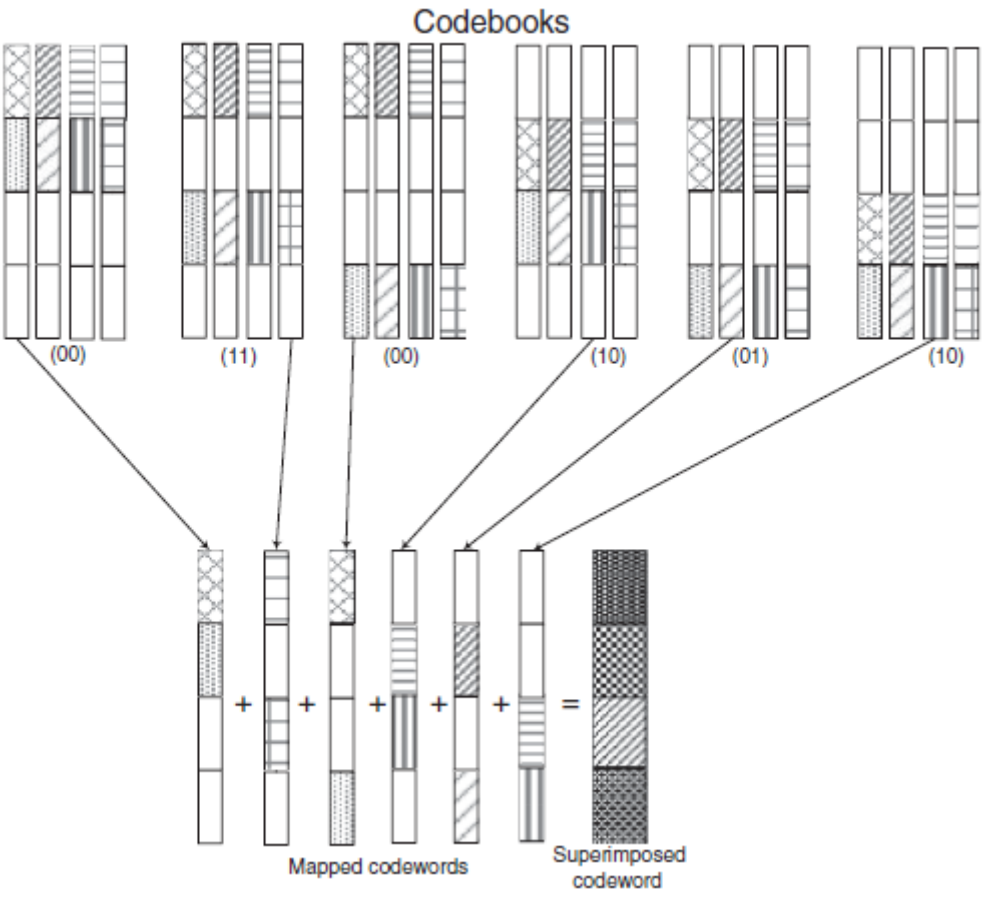
● Dane domeny binarnej są bezpośrednio kodowane do wielowymiarowych słów kodowych domeny zespolonej wybranych z wstępnie zdefiniowanego zestawu książek kodowych

● Wielokrotny dostęp jest osiągany poprzez zaprojektowanie wielu książek kodowych, jednej dla każdej warstwy lub użytkownika

● Słowa kodowe są rozproszone, tak że wykrywanie wielu użytkowników MPA jest stosowalne do wykrywania zmultipleksowanych słów kodowych o przystępnej złożoności

● Transmisja nieortogonalna jest implementowana poprzez multipleksowanie liczby warstw lub użytkowników, która jest większa niż współczynnik rozprzestrzeniania

Bez utraty ogólności możemy założyć, że istnieje K książek kodowych dostępnych dla K użytkowników lub K warstw przestrzennych. Każda książka kodowa składa się z M słów kodowych o długości N, a liczba niezerowych elementów wielowymiarowych w każdym słowie kodowym wynosi d_v. Wszystkie słowa kodowe danej książki kodowej zawierają zera w tych samych wymiarach $N - d_v$, a pozycje zer w różnych książkach kodowych są różne, aby ułatwić unikanie kolizji dowolnej pary użytkowników. W konsekwencji maksymalna liczba książek kodowych jest ograniczona przez wybór N i d_v, równy $\binom{N}{d_v}$. Wartości różne od zera w słowach kodowych mogą przyjmować różne wartości zespolone. Każdy użytkownik lub warstwa mapuje $\log_2 M$ zakodowanych bitów bezpośrednio na wielowymiarowe słowo kodowe zespolone. Wybrane słowa kodowe wszystkich użytkowników lub warstw są nakładane na multipleksowane słowo kodowe, które jest przesyłane przez N współdzielonych jednostek zasobów ortogonalnych, takich jak układy CDMA lub podnośne OFDM. Na przykład, jak pokazano na rysunku,



w systemie SCMA jest K = 6 użytkowników, a każdy użytkownik ma unikalną książkę kodową. Długość słów kodowych wynosi N = 4, a wszystkie słowa kodowe danej książki kodowej zawierają d_v = 2 wartości zespolonych niezerowych w tych samych dwóch wymiarach. Dlatego może obsługiwać do (42) = 6 różnych książek kodowych, a pozycje wartości niezerowych w różnych książkach kodowych są różne, aby ułatwić unikanie kolizji. Dla każdego użytkownika para zakodowanych bitów, np. 00 dla użytkownika 1, 11 dla użytkownika 2, jest mapowana na złożone słowo kodowe w każdej książce kodowej. Wybrane słowa kodowe dla sześciu użytkowników są nakładane na multipleksowane słowo kodowe, które jest następnie przesyłane przez N = 4 współdzielone zasoby ortogonalne

Podsumowanie

Po oryginalnej koncepcji OFDM zaproponowanej przez Changa , nastąpiły dwa fundamentalne przełomy: wydajna implementacja przy użyciu FFT zaproponowana przez Weinsteina i Eberta oraz zastosowanie cyklicznego prefiksu zaproponowane przez Peleda i Ruiza. Następnie OFDM stała się dominującą techniką modulacji dla przewodowych i bezprzewodowych systemów komunikacyjnych w ciągu ostatnich dwóch dekad. Była szeroko stosowana w wielu znanych standardach, xDSL, DVB-T, Wi-Fi, WiMAX, LTE i LTE-Advanced, aby wymienić tylko kilka. Po obszernym porównaniu wszystkich możliwych technik, OFDMA została przyjęta jako jeden z krytycznych elementów 5G NR, jako kompleksowy kompromis między wydajnością, złożonością, porównywalnością i solidnością. Po raz pierwszy w historii komunikacji mobilnej ta sama technika wielokrotnego dostępu staje się podstawą dla dwóch generacji systemów mobilnych. Przewiduje się, że OFDM, OFDMA i SC-FDMA będą również służyć jako krytyczne technologie w nadchodzącej transmisji 6G zarówno w konwencjonalnym paśmie sub-6 GHz, jak i pasmach wysokiej częstotliwości. Ponadto standard 5G zintegrował NOMA ze względu

na swoje zalety w zakresie wydajności widmowej i masywnych połączeń. Przewiduje się również, że zostaną zaprojektowane bardziej zaawansowane i wydajne systemy NOMA oraz ściślej zintegrowane z systemem nowej generacji.