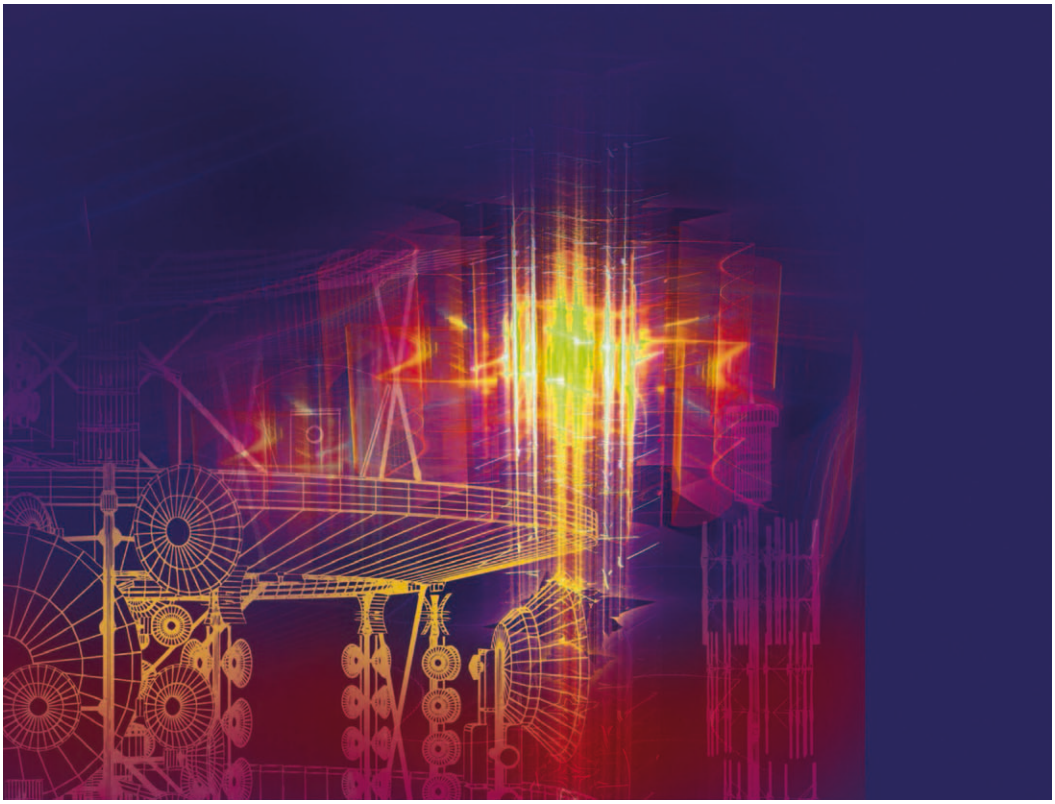
IQ-Kompressionsalgorithmus mit Vivado High-Level Synthese evaluiert

Die Betreiber von Mobilfunknetzen stehen vor großen Herausforderungen, wenn sie bei deren laufender Erweiterung und Verdichtung zufrieden stellende Gewinne erwirtschaften wollen.



Das in diesem Beitrag vorgestellte Kompressionsschema für Funkschnittstellen soll zur Verringerung der Investitionen in die erforderliche Fronthaul-Infrastruktur beitragen. Mit dem Xilinx Vivado HLS Tool zu dessen Evaluierung lassen sich Kosten und Zeitrahmen der Implementierung in den Griff bekommen und die Performance abschätzen.

Wir benutzen das High-Level Synthese (HLS) Tool der Xilinx Vivado Design Suite zur Evaluierung eines Kompressionsschemas im Open Radio Equipment Interface (ORI) Standard für E-UTRA I/Q-Daten zur Abschätzung seiner Auswirkung auf die Signaltreue, die dabei eingeführte Latenz und die Kosten seiner Implementierung. Dabei zeigt sich, dass das Vivado HLS-Tool eine effiziente Plattform zur Evaluierung und Implementierung des gewählten Kompressionsalgorithmus darstellt.

Funknetze unter Druck: mehr Bandbreite!

Der immer weiter steigende Bandbreitenbedarf in Mobilfunknetzen treibt die Forderung nach neuen Netzwerkstrukturen und -fähigkeiten wie MIMO (multiple-input, multiple-output) höherer Ordnung und Carrier-Aggregation. Der damit einhergehende Anstieg der Netzwerk-Komplexität zwingt die Betreiber zu Änderungen in Architekturen, etwa die Zentralisierung des Baseband Processing, um die Ausnutzung der Netzwerk-Ressourcen zu optimieren. Das reduziert die Kosten der Basisband-Verarbeitung, andererseits erhöht die gemeinsame Nutzung dieser Ressourcen die Komplexität des Fronthaul-Netzes.

Diese Fronthaul-Netze, die die modulierten Antennensignale der Carrier zwischen den Basisband-Einheiten (baseband units, BBU) und den entfernten Funkstationen (remote radio heads, RRH) transportieren, sind meist im Common Public Radio Interface (CPRI) Protokoll über Glasfaser implementiert. Das CPRI-Protokoll erfordert eine konstante Bitrate. Seine Spezifikationen wurden über die Jahre hinweg so erweitert, dass seine maximale Datenrate den steigenden Bandbreitenanforderungen gerecht wird. Die Netzwerkbetreiber halten nun nach geeigneten Technologien Ausschau, mit denen sie eine signifikante Erhöhung der Datenrate realisieren können, ohne die Zahl der eingesetzten Glasfasern zu vergrößern. Damit wollen sie den bestehenden Capex- und Opex-Overhead pro Zelle aufrecht erhalten.

Im Sinne von nachhaltigen Langfrist-Lösungen untersuchen die Netzwerkbetreiber außerdem alternative Netzwerkstrukturen. Dazu zählt u. a. der Umbau der Schnittstellen zwischen der Basisbandverarbeitung und den Funkstellen (radio

Autoren
Stefan Petko (stefan.petko@xilinx.com) und Duncan Cockburn (duncan.cockburn@xilinx.com) sind Entwickler bei Xilinx, Inc.

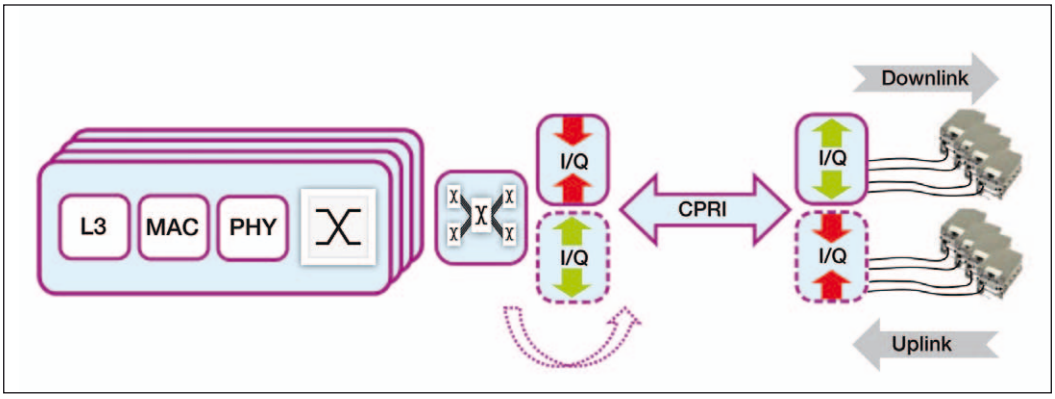


Bild 1: Vereinfachtes Mobilfunksystem mit CPRI IQ-Kompression

units), um die Fronthaul-Bandbreite zu verringern. Allerdings machen es derartige funktionale Umorientierungen unter Umständen eher schwieriger, die stringenten Performance-Anforderungen bestimmter drahtloser Interface-Spezifikationen zu erfüllen.

Eine alternative Methode zur Bandbreitenreduktion besteht in der Implementierung eines Kompressions/Dekompressionsschemas (codec) für Funk-Schnittstellen, die ihren verfügbaren Durchsatz zur Gänze ausschöpfen oder bereits überschreiten. Die dabei erreichbaren Kompressionsraten hängen von der spezifischen Charakteristik der Funksignale ab, wie Rauschpegel, Dynamikbereich und Oversampling-Raten.

Werfen wir einen genaueren Blick auf ein solches Kompressionsschema im ORI-Standard für E-UTRA IQ-Daten – also die realen und imaginären Anteile der übertragenen Modulationssymbole. Eine stark vereinfachte Applikation nach Bild 1 verdeutlicht den Einsatz eines Kompressions- und Dekompressionsmoduls im CPRI IQ Input-Output-Interface.

IQ-Kompressionsalgorithmus

Der ORI-Standard ist eine Verfeinerung der CPRI-Spezifikation zur Realisierung einer offenen BBU/RRH-Schnittstelle. Das neueste ORI-Release spezifiziert eine verlustbehaftete Technik im Zeitbereich zur E-UTRA Datenkompression für Kanal-Bandbreiten von 10, 15 oder 20 MHz. Die Kombination eines 3/4-Resampling mit fester Rate und nichtlinearer Quantisierung von 15-bit IQ-Samples erzielt eine Reduktion von 50 Prozent der Bandbreitenanforderung. Das ermöglicht beispielsweise eine 8 x 8 MIMO-Konfiguration mit Abdeckung von zwei Sektionen über eine einfache CPRI-Verbindung mit 9,8 Gb/s.

Die Resampling-Stufe involviert die Interpolation der I- und Q-Ströme im Eingang, Weiterleitung der interpolierten Daten über ein Tiefpass-Filter und Dezimation des Datenstroms im Ausgang. Der Filterentwurf stützt sich auf die spezifische Kanal-Charakteristik, um die Signalverluste wegen der Downsampling- und Upsampling-Schritte zu minimieren. So erreicht die effektive Bandbreite eines OFDMA-Ausgangssignals in einem 20-MHz E-UTRA-Downlink-Kanal mit 30,72 MHz Sampling den Wert 18,015 MHz - unter der Voraussetzung eines verlustlosen, idealen Tiefpass-Filters mit 3/4-Resampling-Rate.

Der Prozess der nichtlinearen Quantisierung (NLQ) transformiert die 15-bit-Baseband-IQ-Samples mit einer Normalverteilung in quantisierte 10-bit-Werte. Der NLQ-Prozess minimiert den Quantisierungsfehler mithilfe einer kumulativen Verteilungsfunktion (cumulative distribution function, CDF) mit spezifizierter Standardabweichung, um die bei feinerer Granularität häufiger auftretenden Amplituden gegenüber den weniger häufigen stärker zu berücksichtigen. Wie Bild 2b zeigt, füllt die quantisierte Konstellation einen signifikant höheren Anteil ihres geringeren Zahlenbereichs als die in Bild 2a gezeigte Eingangs-Konstellation aus. Dies minimiert den Quantisierungsfehler im Vergleich zu alternativen linearen Quantisierungsverfahren. Die I- und Q-Samples werden meist in einer Lookup-Tabelle implementiert. Sie können entsprechend ihrer jeweiligen Verteilungsfunktionen unabhängig voneinander quantisiert werden.

Vergleichen wir die Performance der ORI-IQ-Kompression mit einer Implementierung eines Mu-Law-Kompressionsalgorithmus nach der ITU-T-Spezifikation G.711. Mu-Law ist ebenfalls eine nichtlineare Quantisierung und verwendet eine logarithmische Funktion zur Redistribution der quantisierten Werte über den verfügbaren Zahlenbereich. Anders als die CDF-basierte Quantisierung, welche die statistische Distribution der Eingangs-Samples berücksichtigt, ist das mit Mu-Law quantisierte Ausgangssignal eine Funktion der korrespondierenden Eingangs-Samples und des spezifizierten Kompander-Wertes.

Um einen Vergleich mit einem äquivalenten Kompressionsverhältnis von 50 Prozent anzustellen, betrachten wir einen 16-auf-8-bit Mu-Law-Encoder. Da dieser kein Resampling erfordert, ist die Mu-Law-Kompression eine kostengünstigere Lösung in Bezug auf Latenz und Implementation. Sie bietet außerdem eine Entscheidungsmöglichkeit zwischen Design-Komplexität und der erzielbaren Signaltreue des rekonstruierten Signals.

Skalierung der Codec-Architektur

Für unsere Prototyp-Konfiguration streben wir die Aufskalierung des Kompressionsalgorithmus an, um einen 9,8304-Gb/s CPRI-Link (Line-Bitrate Option 7) voll auszunutzen. Die laut ORI komprimierte E-UTRA Sample-Spezifikation erlaubt uns die Übertragung von 16 komprimierten IQ-Kanälen (32 unabhängig komprimierte I- und Q-Kanäle) über einen einzigen 9,8G CPRI-Link. Eine Zielvorgabe des Durchsatzes von drei komprimierten Samples pro CPRI-Takt reicht aus, um das 32-bit Xilinx LogiCORE IP CPRI-IQ-Interface voll zu packen. Das ergibt das geforderte komprimierte IP-Ausgangssignal von 737,28 MSamples pro Sekunde.

Mit der Vorgabe eines Domains mit einfachem Takt müssen wir das Resampling-Filter so auslegen, dass es die Ausgangsrate von drei Samples pro Taktzyklus erreicht. Interpolation des Eingangs-Signalstroms mit Nullen erlaubt uns, die nicht zum Signal beitragenden Eingangs-Samples zu vernachlässigen. Der Signalstrom im Ausgang wird somit zu einer Funktion der Interpolationsrate von Subfiltern, wobei jedes Subfilter einen Subset der FIR-Koeffizienten (Gesamtzahl der Koeffizienten/Interpolationsrate) verarbeitet. Jedes der vier parallelen Filter wirkt auf einen Subset von Kanälen. Das hebt den gesamten Durchsatz auf die erforderlichen drei komprimierten Samples pro Taktzyklus an. Zusammen mit dem hohen Durchsatz reduziert diese hier vorgeschlagene Architektur auch die Latenz des Resampling, da in jedem Subfilter nur ein Teil der Koeffizienten genutzt wird.

Im Kompressionspfad berechnen wir die NLQ-Quantisierungstabelle mithilfe der kumulativen Verteilungsfunktion (CDF). Unter der Annahme einer

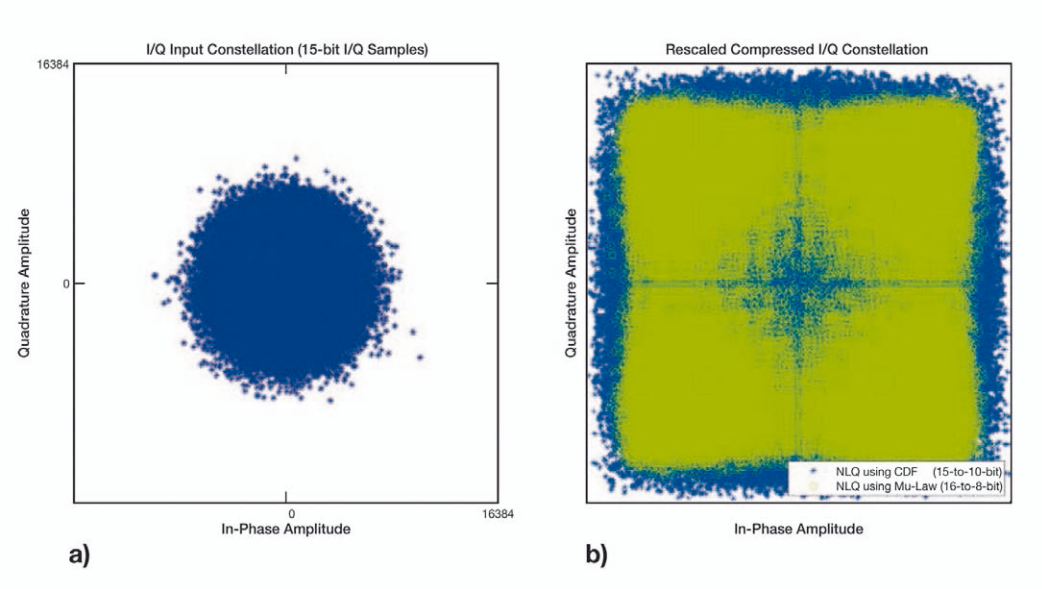


Bild 2: IQ-Anordnung des Referenz 20-MHz E-UTRA DL Kanal-Eingangs-Frame (a), und die komprimierten IQ-Daten (skaliert) zur Illustration des effektiv nutzbaren Zahlenbereichs für jede Konstellation (b)

symmetrischen IQ-Distribution verkleinern wir die NLQ-Lookup-Tabelle auf 214 Einträge mit auf 9-bit quantisierten Werten. Da unser Design drei parallele Lookup-Vorgänge pro Zyklus erfordert, implementieren wir drei parallele Lookup-Tabellen mit denselben Quantisierungswerten. Die Quantisierungsebenen können für die I- und Q-Samples unabhängig berechnet werden, wozu man die erwarteten oder beobachteten Werte der Standard-Abweichung heranzieht. Alternativ können auch Subsets von Kanälen unabhängig voneinander auf der Basis der aktuellen Signalpegelmessungen oder von Netzwerkparametern höherer Ordnung quantisiert werden. Zur Dekompression verwenden wir die Quantile-Funktion (Inversion von CDF), um die inverse NLQ-Tabelle zu berechnen. Deren Größe ist auf 29 Einträge von 14-bit-Werten limitiert.

Wir haben den implementierten Codec-Algorithmus mit Anregung durch einen 20-MHz LTE E-UTRA FDD-Kanal-Stimulus getestet, der mit der MATLAB LTE System-Toolbox generiert wurde. Anschließend setzten wir den Keysight VSA zur Demodulation der erfassten IQ-Daten ein und quantifizierten die Signalverzerrung infolge der Kompressions- und Dekompressionsschritte durch die Messung der Größe des Fehlervektors (error vector magnitude, EVM) in der Ausgangs-Wellenform. Wir haben die erhaltenen EVM-Messungen im Ausgang – die den Unterschied zwischen dem idealen und dem gemessenen Signal repräsentieren – mit dem Referenz-Eingangssignal EVM verglichen.

High-Level Modellierung und Implementierung

Wir begannen die Implementierung mit der Entwicklung von einkanaligen Kompressions- und Dekompressionsmodellen in der GNU Octave Sprache, unter Nutzung der Packages für Signalverarbeitung und Statistik. Zusätzlich zu den sehr nützlichen Datenpunkten für die Verifikations-Referenz erzeugt das Modell auch einen Satz von FIR-Filterkoeffizienten und Quantisierungstabellen. Mit dem Ansatz eines anspruchsvollen mathematischen Modells bietet das Vivado HLS-Tool einen klaren Übergangspfad zur Evaluierung der vorgesehen Architektur in Bezug auf die potenzielle Performance der Hardware und deren Kosten. Wir haben unsere C++-Testbench so eingerichtet, dass sie die Eingangs-Datenströme mit der Komprimierungs- und Dekomprimierungsfunktion bearbeitet. Wir synthetisieren diese Funktionen unabhängig voneinander, da sie physisch an den entgegen gesetzten Enden eines CPRI-Link platziert sind. Wir implementierten alle externen und internen funktionalen Interfaces unter Nutzung der HLS-Ströme mit dem Interleave-Datenfluss des Kanals über einfache C++-Loops.

Wir setzten die Vivado HLS FIR IP zur Erstellung des Prototyps für das Resampling-Filter ein. Um die hohen Anforderungen unseres Designs im Durchsatz zu erfüllen, implementierten wir parallele Single-Rate FIR-Filter und verwenden eine Dezimation mit Loop-basiertem Filterausgang.

Es ist natürlich möglich, ein Resampling-Filter mit besserer Ressourcen-Effizienz zu erstellen, indem man es als Polyphase-Filter implementiert. Eine derartige, einfach einsetzbare Alternative mit Unterstützung der ORI-Resampling-Raten ist das Mehrkanal-Filter mit verringerter Sample-Rate, wie es in der Xilinx Application Note XAPP1236 (Multi-Channel Fractional Sample Rate-Conversion Filter Design Using Vivado High-Level Synthesis) beschrieben wird.

Die Vorteile einer schnellen C-Simulation werden noch klarer, wenn die zu verifizierenden Datensätze eine gewisse Größe ereichen. Das ist der Fall, wenn man einen IQ-Kompressions-Algorithmus evaluiert. Denn ein voller Datenrahmen (307.200 IQ-Samples pro Kanal) ist als Minimalwert erforderlich, um die VSA-Tools für EVM-Messungen einzusetzen. Wir haben eine um zwei Größenordnungen höhere Geschwindigkeit für die C-Simulation erzielt, verglichen mit der C/RTL Ko-Simulation. Das macht aus der neunstündigen Ko-Simulation eine fünfminütige C-Simulation für unseren IP-Test-Run der Kompression.

Ein weiterer wichtiger Vorteil beim Einsatz der HLS-Testbench ist die Leichtigkeit der Nutzung der Eingangsdaten und der Erfassung der Ausgangsdaten auf der Basis der HLS-Ströme. Das Ergebnis war die Bereitstellung eines Interface für die Datenanalyse mit VSA-Tools oder ein direkter Vergleich mit dem Ausgang des Octave-Modells auf der C++-Testbench.

Performance-Messungen

Die Messungen mit dem Keysight VSA ergaben einen mittleren EVM-Wert von 0,29 Prozent für eine Codec-Konfiguration mit 144 FIR-Koeffizienten. Im Vergleich mit den Original-Eingangsdaten mit einem EVM RMS-Wert von 0,18 Prozent beträgt das zusätzliche, der Kompressions-Dekompressionskette zuzuschreibende EVM 0,23 Prozent. Verglichen mit dem Mu-Law-Kompressionsalgorithmus resultiert das bei äquivalenten Eingangs-Datensätzen in einem mittleren EVM von 1,07 Prozent.

Verringerte Latenz und kostengünstigere Ressourcennutzung machen die Mu-Law-Kompression zu einer besseren Wahl gegenüber der von ORI vorgeschlagenen IQ-Kompression. Das gilt für Situationen, in denen man mindestens 1 Prozent des

EVM-Budgets (von 8 Prozent der gesamten LTE-Downlink-Signalverarbeitung) der IQ-Kompression zuweisen kann. Allerdings impliziert jedwede zusätzliche Signalverzerrung engere Performance-Ziele für die verbleibenden Systemkomponenten. Die potenziellen Kostenvorteile der IQ-Kompression werden also eventuell durch die Kostensteigerungen der digitalen Frontend-Komponenten und der Leistungsverstärker wieder aufgelöst.

Die Vivado High-level Synthese bestätigte den geforderten Durchsatz in Bezug auf das Initiationsintervall – also die Zahl der Taktzyklen, bevor der Top-level Task neue Eingangsdaten akzeptieren kann. Wir haben außerdem verifiziert, dass die exportierten Vivado IP Integrator-Kerne die Timing-Anforderungen der Kintex UltraScale-Plattform erfüllen.

Wir haben unsere Untersuchung hier auf einen kleinen Teil möglicher Konfigurationen und Vektoren von Eingangsdaten beschränkt. Wenn jedoch das System-Modell und das zugehörige C-Modell einmal erstellt sind, ist es möglich, auch alternative Konfigurationen in wenigen Minuten zu spezifizieren, zu implementieren und zu evaluieren.

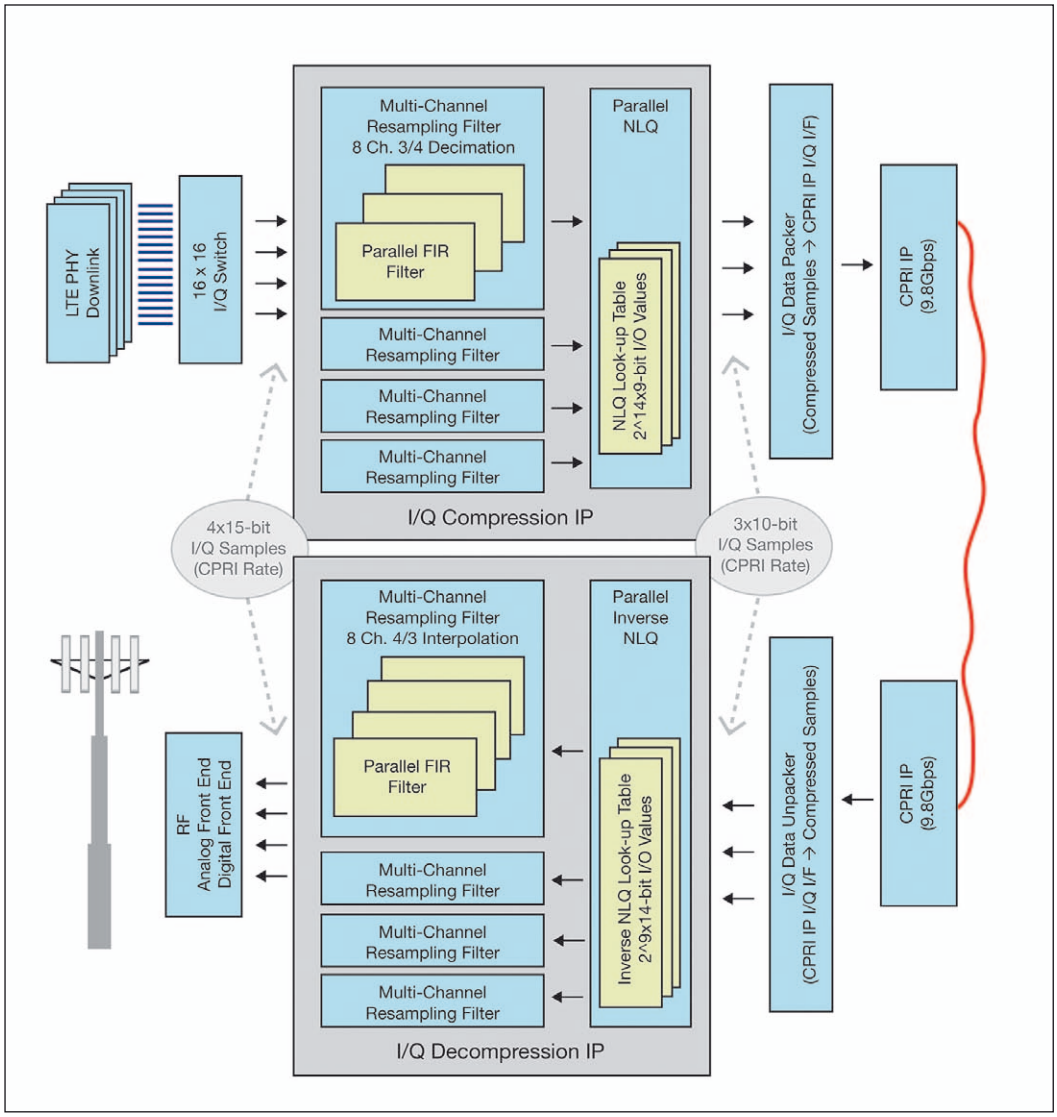


Bild 3: IQ-Codec-Architektur mit den Verarbeitungsraten der Samples an den IP-Schnittstellen des Codec (dargestellt ist nur der Downlink-Pfad)

Design-Alternativen

Aus der Design-Tool Perspektive bietet Vivado HLS einen sehr vorteilhaften Pfad zum Hardware-Prototyping. Eine High-level Testbench passt sehr gut in ein Design-Framework, das einen Datenfluss innerhalb mehrerer Design- Und Verifizierungs-Tools erfordert. Der Hauptvorteil einer solchen Testbank liegt in der Möglichkeit schneller C-level Simulationen eines Hardware-Systemmodells. Zur IQ-Kompression und ähnliche Applikationen involvieren Simulationen häufige Änderungen der Parameter auf höheren Ebenen oder Änderungen der Eingangs-Datensätze. Alles das macht ein schnelles Feedback zu einem essenziellen Vorteil.

Messungen zeigen, dass das vorgeschlagene ORI-Kompressionsverfahren Signalverzerrungen von weniger als 0,25 Prozent in einem 20-MHz E-UTRA Downlink-Kanal liefern kann. Während die Kompressions-Performance von der Charakteristik des Eingangskanals abhängt, bietet diese ORI-Kompression auch eine Möglichkeit zur Performance-Einstellung, und zwar durch die Wahl des optimalen Filter-Designs und der Quantisierungs-Parameter.

Unser Prototyp nutzt einen gemeinsamen statischen Satz von Design-Parametern für alle 16 Antennen-Carrier-Datenströme. In einem praktisch realisierten System ist die Signal-Charakteristik entweder im Vorhinein bekannt, oder sie lässt sich messen und zur Abstimmung des Systems nutzen. Alternativ könnte die Kompressions-Performance dynamisch durch Rekonfiguration der Quantisierungstabellen eingestellt werden, um die geforderte Mindest-Signaltreue zu erhalten.

Die Kosten in Bezug auf die implementierten Ressourcen und die zusätzlich notwendige Latenz, um die Kompression und Dekompression auszuführen, sind ebenfalls im Zusammenhang mit der geforderten Kompressions-Performance zu betrachten. Die Größe des Resampling-Filters und dessen Latenz bestimmen die Gesamtkosten der Codecs. Eine größere EVM-Toleranz ermöglicht Designs mit weniger Filterkoeffizienten.

Fazit

Unter dem Gesichtspunkt der erzielbaren Time-to-Market hat Xilinx eine Konzeptstudie für einen ORI-basierten IQ-Codec erstellt. Weitere Informationen und Zugriff auf die Design-Dateien finden sich auf der Xilinx Website unter www.xilinx.com/applications/wireless-communications/wireless-connectivity.html, oder sind per E-Mail an Perminder Tumber, Wireless Business Marketing Manager für Connectivity: Permind@xilinx.com, verfügbar. ◄