5G Massive MIMO Innovations

Boosting Spectral, Energy and Site Efficiency

White paper
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive Summary</td>
<td>3</td>
</tr>
<tr>
<td>Introduction</td>
<td>4</td>
</tr>
<tr>
<td>Spectral Efficiency</td>
<td>5</td>
</tr>
<tr>
<td>Advanced Beamforming</td>
<td>6</td>
</tr>
<tr>
<td>Multi-User Precoding</td>
<td>7</td>
</tr>
<tr>
<td>Distributed Massive MIMO</td>
<td>11</td>
</tr>
<tr>
<td>Energy Efficiency</td>
<td>13</td>
</tr>
<tr>
<td>Adapting Array Configuration for Energy Saving</td>
<td>13</td>
</tr>
<tr>
<td>Site Efficiency</td>
<td>16</td>
</tr>
<tr>
<td>Millimeter Wave Massive MIMO on Existing Sites</td>
<td>16</td>
</tr>
<tr>
<td>Massive MIMO in Rural Deployments</td>
<td>17</td>
</tr>
<tr>
<td>Summary and Outlook</td>
<td>21</td>
</tr>
<tr>
<td>Appendix</td>
<td>22</td>
</tr>
<tr>
<td>A1: Details on NR type II CSI and ETF</td>
<td>22</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>23</td>
</tr>
<tr>
<td>Further Reading</td>
<td>24</td>
</tr>
</tbody>
</table>
Executive Summary

Massive MIMO (Multiple Input Multiple Output) is an important building block and integral part in 5G systems, as a large number of antennas allows for highly directional transmission via beamforming. By ensuring adequate distance between antennas, numerous users can be served by the same time-frequency resource, boosting spectral efficiency.

This paper discusses innovative evolution options for massive MIMO and shows the potential for improving spectral and energy efficiency. Despite a decade of mMIMO research, there is still no ultimate solution that manages to balance complexity, energy consumption and spectral efficiency. Bell Labs has an active role in creating mMIMO innovations.

Spectral efficiency of massive MIMO can be boosted by innovative beamforming algorithms - more complex algorithms yield higher gains. The paper shows that a first phase beamforming solution can already provide excellent performance by using the improved grid-of beams concept, high-dimensional codebooks and especially by using Eigenbeamforming, which gives a particular boost to cell edge performance. Further improvement can be provided by multiuser MIMO precoding using instantaneous channel knowledge. In an urban micro scenario, for example, algorithms such as robust zero-forcing provide roughly a further 30% user rate gain compared to Eigenbeamforming, while not requiring a perfect knowledge of the channel. An additional performance boost can be provided by distributed MIMO, offering a potential further 50% gain in spectral efficiency compared to Eigenbeamforming.

Reducing the power consumption of massive MIMO systems is important in cutting OPEX and saving CO₂ emissions. The adaptation of the array configuration by load-dependent switching-off of transceiver chains can bring down mMIMO power consumption by 75%, while keeping the average user rate constant (provided that significantly fewer users are active). For a typical day-night-cycle of varying load, this means a 30% reduction in power consumption.

Massive MIMO also provides better coverage and throughput in rural deployments, even when fewer simultaneously active users are present. As seen in both urban and rural deployments, larger arrays are more future proof and can significantly delay the time when the system will need upgrading to deal with increased data demands. Large scale arrays coupled with a 5G air interface can service much larger cell sizes due to 5G's ability to use beamforming for common control channels.

For offloading traffic to higher mmWave bands, the same sites can be shared for both sub-6GHz as well as e.g. 28 GHz. With a 200m inter-site distance, for example as seen in cities, more than 10 bps/Hz of traffic can be carried per cell by a mmWave system, freeing up the lower band.

Nokia is an experienced partner in massive MIMO technology, deploying massive MIMO globally in both LTE and 5G networks in millimeter wave and sub-6 GHz bands in TDD and FDD networks in all key global markets. Nokia has introduced powerful and energy efficient ReefShark chipsets into massive MIMO products to optimize cost and power consumption.
Introduction

An attractive solution for boosting mobile network performance, beamforming can provide higher spectral efficiency, bringing a lot more capacity to existing base station sites. Beamforming can also improve link performance and provide extended coverage. Its benefits are obtained in practice with massive MIMO (Multiple Input Multiple Output) antennas. 5G radio is designed to support massive MIMO. The underlying principle of beamforming is illustrated in Figure 1. The traditional solution transmits data over the whole cell area while beamforming sends the data to users via narrow beams. The same resources can be reused for numerous users within a sector, while the interference can be minimized and the cell capacity maximized.

Figure 1. Beamforming with massive MIMO enhances radio capacity and coverage

Massive MIMO is part of some 4G networks and most 5G deployments, particularly on TDD bands between 2.3 and 4.9 GHz. Nokia Bell Labs aims to make massive MIMO even better via a number of innovations. This paper illustrates several solutions that will produce a further boost to the spectral efficiency, energy efficiency and site reuse efficiency of massive MIMO. Spectral efficiency is important for maximizing throughput with a given spectrum, energy efficiency is important for minimizing the power consumption and site efficiency is important to maximize the value of existing base station sites by using all of the 5G spectrum.

Figure 2. Targets of Nokia 5G massive MIMO innovations

The basic principle of MIMO processing originates back to schemes such as BLAST (Bell Labs Layered Space-Time), invented in Bell Labs by Gerard Foschini et al [F96]. A key role in the invention of the Massive MIMO concept was played by Thomas Marzetta et al. [M10] in Bell Labs.
Spectral Efficiency

Beamforming and spatial multiplexing (i.e., MIMO) methods play a fundamental role in 5G NR with its beam-centric design. Radio units with large antenna arrays allow the real-time/user-specific adaptation of the radiation pattern. With beamforming, the radiation pattern can be directed to a user so that the link signal-to-interference-plus-noise ratio (SINR) is maximized. Spatial multiplexing implies the concurrent transmission/reception of multiple data layers (to a single or to multiple users), each layer with a dedicated radiation pattern chosen such that the SINRs of individual layers optimize a particular metric, for example, data rate per user.

Early 5G NR systems for the sub-6GHz frequency range 1 (FR1) used only a small number of predefined beams for data transmission (e.g., the same beams used for the Synchronization Signal/PBCH block (SSB) transmission), establishing good coverage and robust eMBB functionality.

The implementation of advanced beamforming and precoding features for FR1 can be split into three innovation and complexity categories as illustrated in Figure 3. The achievable spectral efficiencies are shown in Figure 4.

Figure 3. Illustration of the advanced beamforming (Phase I), multi-user precoding (Phase II), and distributed precoding (Phase III) methods for massive MIMO
In the following, we summarize the key innovations of phase I and II and provide an outlook to phase III.

### Advanced Beamforming

In phase I, refined/UE-specific beams for higher beamforming gain and better user separation are implemented. This improves the cell edge performance and establishes robust multi-user MIMO performance gains. There are three beamforming methods used to accomplish this goal:

a. **Grid-of-Beams (GoB):** A pre-defined grid of coverage- and sidelobe-optimized beams is used, which comprises, for example, 8 SSB beams and 32 refined beams. For beam management, the gNB transmits CSI-RS that are used by the UEs to determine and report the strongest or two strongest beams. Alternatively, uplink (UL) measurements (e.g., sounding reference signals (SRS)-based) can be used by the gNB to select the best beams for each UE. Channel measurement and data transmission are carried out on a low-dimensional/UE-specific channel (with e.g., 4 antenna ports), that is established by the selected beams.

b. **Codebook-based Beamforming:** A very small number of beams (e.g., one sector beam, or four quarter-sector beams, each transmitted on sub-panels) is used to create a high-dimensional channel with up to 32 antenna ports. In the case of one sector beam, no beam management is needed. The CSI-RS transmission for channel measurement is used by all UEs in the cell to select the high-dimensional codeword (i.e., beamformer for data transmission) that yields the highest throughput.

c. **Eigenbeamforming (EBF):** SRS-based uplink measurements are exploited by the gNB to estimate the dominant Eigendirections of the propagation channel, and are used as beamformers (so-called

---

**Figure 4.** Downlink spectral efficiencies of the phase I to III beamforming/precoding methods. Simulation scenario: 3GPP 3D-UMi (200m ISD), 3.5 GHz carrier frequency, 100 MHz bandwidth, DDDSU slot structure, 64 TXRUs per cell, users with 2T4R, SU-MIMO or MU-MIMO with proportional-fair user scheduling with maximum of four co-scheduled users, SRS every 10msec on 16 PRBs with frequency hopping.
Eigenbeamformer) for transmitting/receiving data on a low-dimensional (UE-specific) channel (with e.g., 4 antenna ports). In contrast to the GoB- or codebook-based methods, a single Eigenbeam can transmit/receive along multiple dominant channel directions and thereby increase the beamforming gain. Such a method is particularly well suited for time division duplex (TDD) systems where uplink and downlink use the same frequency. For the DL beamforming in frequency division duplexing (FDD) systems, an additional transformation step for the Eigenbeams is needed to compensate the frequency difference between uplink and downlink, changing the antenna element distance relative to the wavelength.

The benefit of the three beamforming methods is illustrated in Figure 4 for the downlink of an Urban Micro cellular network:

- For single-user (SU) MIMO, the performance gains with respect to the SSB beamforming are particularly large at the cell edge. Those gains come from the increased beamforming gain (while the average interference power remains unchanged) that improves the receive SINR and thus the user rates. Further, codebook-based beamforming outperforms the GoB-based method. This is because the UEs select beams that maximize the SU-MIMO throughput instead of the received signal power as is the case in the beam management of the GoB-based method. Replacing the GoB-based method by Eigenbeamforming, however, can provide additional performance gains because the Eigenbeams are able to adapt to and exploit the characteristics of the rich-scattering environment in the Urban Micro channel.

- For the multi-user (MU) MIMO setting, the refined/user-specific beams reduce the (cross-beam) interference between the users within the cell, so that the scheduler can pair more users (i.e., data layers) and establish good spatial multiplexing gains. Important to note is the performance difference between the GoB- and codebook-based methods. Here, the reduced sidelobe design of the GoB beams reduces the co-channel interference and thus yields a higher average user rate.

Other aspects that affect the choice between the GoB- and codebook-based methods are the CSI-RS overhead and coverage. With a GoB, the UE-specific CSI-RS for channel measurement is sent on narrow/high-gain beams which provide good CSI-RS coverage at the cell edge but can imply a larger CSI-RS overhead. For the codebook method, in contrast, cell-specific CSI-RS is sent on a smaller number of (broad) sector beams, whose reception at the cell edge can be challenging. The choice will thus depend on the deployment scenario and traffic conditions.

For the Eigenbeamforming method, a fast beam adaptation can be challenging with large numbers of active users where the limited SRS resources must be shared. Also, mobile users at the cell edge with limited Tx power must use narrowband SRS with frequency hopping to increase their power spectral density, which implies a slower beam adaption. Therefore, SRS-based beam management can be applied only to a limited set of eligible users. For those users, it provides a faster beam adaptivity and higher rank of beamformed channel in comparison to the GoB-based method.

**Multi-User Precoding**

The objective of phase II (downlink) precoding methods is to use enhanced spatial multiplexing (i.e., higher number of layers) without sacrificing the baseline (particularly the cell edge) performance established in phase I. By incorporating the interference between co-scheduled users (so-called intra-cell interference) into the precoder design, multiple users can be simultaneously served with high link SINRs and thus improved data rates. The only residual performance degradation arises from the inter-cell interference. Since there is no need to mitigate the interference between co-scheduled users by using sharp beams and appropriate user scheduling, the multi-user precoder operates directly on the transmit/receive chains of the antenna array (instead of antenna ports created by beams).
These advanced precoding methods firstly focus on TDD systems (with calibrated antenna arrays) where explicit channel state information (CSI) can be acquired in the UL through SRS transmissions. FDD systems will follow once there is enough penetration with UEs that support explicit (DL) CSI feedback (i.e., enhanced Type II CSI feedback as described in the “CSI feedback optimization” section). One should note that in UL measurement-based methods, the maximum channel rank per user is limited by the number of Tx antennas that can be used for SRS transmission. For that reason, Type II CSI feedback methods can also play an important role for TDD systems.

There exists a large variety of (linear) precoding methods (e.g., regularized zero-forcing, maximum signal-to-leakage-and-noise-ratio, block-diagonalization) whose performance and required processing power can differ substantially. The optimal algorithm design is robust with respect to imperfect CSI (e.g., due to channel aging and UE Tx power limitations) and minimizes the number of calculations that need to be carried out at the scheduler’s time scale. In Figure 5, for example, we compare the block-diagonalization (BD) method against a robust zero-forcing (ZF) method for various SRS sub band sizes. As an upper performance bound, we include the performance achieved with perfect channel state information (i.e., genie CSI).

As expected, the computationally much more complex BD method outperforms the robust ZF method when perfect CSI is available at the gNB. If the CSI is acquired through sub band SRS then the performance of the BD method deteriorates severely due to the CSI aging effect. Furthermore, one can observe a trade-off between the CSI aging effect and the UE Tx power limitation. Large SRS bandwidths allow a more frequent measurement of the channel but are not applicable for cell edge UEs due to the Tx power limitation. In contrast, the robust ZF method copes well with the CSI aging effect, and thus can avoid large SRS bandwidths.

Figure 5. Downlink performance of phase II multi-user precoding methods, comparing a robust zero-forcing variant against the block-diagonalization method, as a function of SRS sub band size. Simulation scenario: 3GPP 3D-UMi (200m ISD), 30 km/h user speed, 3.5 GHz carrier frequency, 100 MHz bandwidth, DDDSU slot structure, 64 TXRUs per cell, users with 2T4R, proportional-fair user scheduling with max. 4 co-scheduled users, precoding based on UL CSI with maximum rank two per user, SRS every 10 msec with various sub band sizes and frequency.
CSI feedback optimization

Massive MIMO performance depends on the accuracy of the CSI. It is possible to obtain CSI based on uplink measurements and based on feedback from UE. In FDD systems, due to the absence of channel reciprocity, the UE must feed back the DL CSI to the gNB.

This section shows the tradeoff between CSI reporting and downlink capacity. The general trend is that more frequent CSI reporting can improve downlink performance while consuming a part of uplink capacity. Moreover, explicit CSI feedback is needed to support advanced MIMO concepts envisaged in upcoming 3GPP releases, such as non-linear precoding and coherent joint transmission multi-TRP precoding. This section shows that new uplink CSI schemes can improve downlink performance while minimizing uplink overhead.

In Release 15, NR type II CSI feedback is promising significant performance gain over legacy LTE codebooks, more than 30% throughput enhancement in some cases [V18]. However, this gain comes at the expense of significant increase in UL overhead [R1-17]. Further details on NR type II CSI in Rel.15 can be found in Appendix A1.

Enhancement of type II CSI feedback for Rel. 16 was agreed in 3GPP based on exploiting the frequency correlation of the communication channel to reduce the UL CSI overhead by a DFT based frequency domain (FD) compression scheme. [R1-19]. Further details on NR type II CSI in Rel.16 can be found in Appendix A1.

In this section, we compare explicit time domain compression using Orthogonal Matched Pursuit (OMP) compressive sensing (Explicit Time-domain Feedback, ETF) [V18, R1-17] against state-of-the-art Release 16 NR type II CSI.

Similarly, ETF as discussed in [AVW18, AJW19], exploits the time domain channel sparsity to reduce the UL CSI overhead by applying a projection matrix derived from an IDFT codebook. Other than the fact that a compressive sensing scheme was used in [AVW18, AJW19], the main difference between ETF and NR type II CSI in Rel.16 would be the input signal to the DFT/IDFT compression block. In the case of NR type II CSI Rel.16, the input signal is a matrix of approximated eigenvectors, and the final output is the precoder matrix indicator (PMI) per layer. In the case of ETF, the input is the ‘raw’ channel frequency matrix (hence CSI is “explicit”) and the output is the time domain channel impulse response per UE receive antenna.

In Figures 6 and 7, the normalized user performance throughput (UPT) is depicted vs the required UL overhead for three CSI feedback schemes: Rel.15 NR type II CSI, Rel.16 NR type II CSI and ETF. We used a 16-element antenna array in a DUma channel with 10 MHz bandwidth and 4 GHz carrier frequency. The UE had two receive antennas. A bursty traffic system was used with resource utilization of 50%.

Both NR type II CSI Rel.16 and ETF used L=4, where L is the number of spatial domain (SD) beams per polarization. For NR type II CSI Rel.16, we used M=7 coefficients per layer and the maximum number of non-zero FD coefficients fed back (regarding the time-domain compression exploiting sparsity). For ETF, M=4 and the maximum number of non-zero FD coefficients is fed back. (For further details on L and M see appendix A1.)
As we can see, for MR1 NR type II CSI Rel.16 provides the best performance-overhead trade-off. The comparison is different for the MR2 case, where the overhead of the ETF doesn't change, since it is still compressing the same information on the channel frequency response. On the other hand, the overhead of NR type II CSI Rel.16 is higher since the PMI must be compressed on two layers for users who chose rank=2.
Distributed Massive MIMO

In a conventional cellular network, each user is served by a single cell at a time on a given frequency carrier, and transmissions to/from a user are experienced as interference by co-channel users in other cells. This inter-cell interference imposes a fundamental limit on the achievable spectral efficiency. While the existing higher-layer-based multi-connectivity features of LTE and NR allow also for connection to multiple sites (e.g. with carrier aggregation allowing for inter-RAT operation, increasing robustness and improving mobility) they will not directly deal with the inter-cell interference problem. Physical-layer-based multi-transmission-and-reception-point (multi-TRP) signal processing will address this problem, especially with coherent multi-antenna multi-site processing.

The idea behind distributed MIMO (D-MIMO) is to suppress inter-cell interference (actually turning it into useful signal contributions) through joint and coordinated beamforming at the antenna arrays of multiple cells towards a set of mutually interfering users, thereby boosting the spectral efficiency of the network. Coordinated beamforming across cells can be achieved with centralized baseband processing, which facilitates the necessary exchange of signal and channel state information between cells.

It should be noted that the D-MIMO concept was studied extensively in the days of 4G LTE standardization, under the name Cooperative Multi-Point (CoMP). Uplink LTE CoMP has been successfully used to increase uplink capacity in mass events and in busy areas by a factor of three using Nokia’s Centralized RAN solution.

The theoretical and simulation studies demonstrated the potential for 50% or more spectral efficiency gain also in downlink, but CoMP did not gain a foothold in actual practice, the main reasons being the following: (a) the backhaul bandwidth and latency requirements seemed too daunting at the time; (b) the most pressing need for higher spectral efficiency was on the downlink of FDD networks, unfortunately the least favorable scenario for D-MIMO (owing to the difficulty in acquiring accurate channel state information at the network side); and (c) limitations on baseband processing power meant that sophisticated channel estimation and beamforming algorithms could not be readily implemented.

A whole generation of technology later, it could be argued that the above limitations no longer hold (or, at least, do not hold to the same degree), and therefore that D-MIMO deserves another close look. Also worth mentioning here are the improved CSI feedback mechanisms in 5G (described in earlier sections), which make D-MIMO potentially attractive even for the FDD downlink. However, in the era of massive MIMO, some new considerations emerge. When a cellular network is equipped with massive MIMO antenna arrays, each cell has the ability to suppress inter-cell interference autonomously (without the need for joint beamforming across cells), suggesting that D-MIMO would be superfluous.

However, suppressing inter-cell interference in this manner must come at the expense of the ability to suppress intra-cell interference (which underlies the spatial multiplexing gains due to massive MIMO). It is therefore important to understand this tradeoff quantitatively, and to determine if D-MIMO can still offer a significant spectral efficiency boost in a massive MIMO network. It is also important to study the robustness of D-MIMO gains for imperfect CSI (arguably the dominant impairment in practice) in such a network, taking into account issues such as pilot contamination.

Recent research at Bell Labs has addressed these questions and shown that D-MIMO has the potential for around a 50% spectral efficiency gain even in massive-MIMO-enabled networks (at least up to 64 antenna ports per cell, assuming four users per cell). However, achieving these gains requires large enough clusters of coordinating cells, ideally user-specific (e.g., all cells with signal strength within 20 dB of the primary serving cell), as well as sophisticated channel estimation algorithms that can fully exploit channel coherence in time, frequency, and space.
Further, it has been demonstrated that the beamforming algorithms can be made more robust to withstand CSI imperfections by incorporating the CSI error statistics into the computation of beamforming weights and transmit powers. Finally, a new beamforming approach inspired by uplink-downlink duality has been developed, which outperforms conventional techniques like regularized zero-forcing (RZF) and maximum ratio transmission (MRT) and is suitable even for smaller clusters of coordinating cells (unlike RZF).

The figure below shows some exemplary results on potential D-MIMO gains on the TDD downlink (with CSI acquired from uplink reference signals), with either four or 64 antenna ports per cell. Each cell serves four users, and each user has a single antenna. The D-MIMO cell cluster serving a user comprises all cells with signal strength within 20 dB of the primary serving cell (as a result, the average user is served by ~5 cells).

The channel model is Urban Macro (as in 3GPP 38.901), with a 500 m inter-site distance, and the carrier frequency is 3.5 GHz. The main takeaway here is that the gain in median user rate due to D-MIMO is virtually the same (about 55%) both for four ports/cell and for 64 ports/cell, and therefore that D-MIMO remains a promising technique for enhancing spectral efficiency even in a massive MIMO network. The findings from figure 8 answer the question about whether more antenna elements make multi-cell coordination via D-MIMO less valuable and the answer is no.

Figure 8. D-MIMO gains in DL with 4 ports/cell (left) and 64 ports/cell (right) compared to single cell only transmission. Each user is served on average by 5 cells.

Relative gains with distributed MIMO
Energy Efficiency

Adapting Array Configuration for Energy Saving

5G massive MIMO systems are designed to provide a significantly improved system throughput and service quality. This is achieved by spatial multiplexing using a system design with an increased number of MIMO layers, as well as an increased bandwidth as defined by the 3GPP NR standard. The energy consumption is determined by the hardware driving a high number of antenna ports and antenna array elements. However, in varying load conditions, when the number of simultaneous UEs (simultaneous MIMO layers) is low, the achievable throughput is also much lower, despite the immense hardware effort and related energy consumption at the base station. Energy efficiency is an important parameter, which affects the cost per bit, and therefore high energy consumption unnecessarily increases operational costs.

To overcome this reduction in energy efficiency at low load conditions, we can adapt the antenna array usage, while keeping the average throughput per UE the same. Different energy saving schemes can be applied depending on the architecture of the array. An example architecture is shown in Figure 9. Since the achievable throughput depends on the number of antenna ports per MIMO layer, it is reasonable to switch off complete antenna ports to save energy at times of low load. This comprises the complete data and RF conversion units, transmit chain with PAs and receive chain with LNAs, filters and gain blocks. In the architecture in Figure 9, a complete subpanel can be switched off by deactivating one of the P antenna ports. Alternatively, only some of the elements per subpanel can be switched off, while keeping the number of antenna ports fixed. This maintains the degrees of freedom for the MIMO layers, while saving energy by reducing the total transmit power per subpanel.

Figure 9. Basic hybrid array architecture, showing different switching-off possibilities in the processing chain, either at the full antenna port and RF chain level or within each antenna port deactivating subarray elements.
Figure 10 shows the shapes of the adapted hybrid arrays for each type of complete port switching. Array type “A” in Figure 10 consists of 256 elements, where every four vertically grouped cross-polarized elements represent one subpanel. Each polarization is related to one physical antenna port, with a total of 64 antenna ports. Reducing the number of ports by factors of 2 and 4 leads to types K and L in Figure 10, with 128 and 64 elements, respectively.

The average UE throughput for the different array types is shown on the right of Figure 10. When the number of simultaneously served UEs decreases from 12 to eight and four, respectively, the average UE throughput will remain with the adapted array size of type K and L, as indicated by the points a, b and c. More information can be found in [HWW+18]. (Note that there may be a remaining coverage challenge in coverage-limited, thus non-interference-limited areas.)

To assess how much energy is saved by the different sized arrays, a simple energy consumption model has been derived. It comprises transmit and receive side conversion units including digital-to-analog and analog-to-digital conversion and RF mixers, PAs on the transmit and LNAs on the receive side. It also comprises filters, splitters and combiners which affect the required PA output power. The energy saved by the array adaptation schemes depends on the architecture. In the example considered above, the number of transmit and receive conversion units increases with the number of antenna ports, whereas the number of PAs and LNAs increases with the number of antenna elements.

The analysis of energy consumption analysis shows that energy consumption is strongly reduced when using the smaller arrays. In Figure 11, the relative power reduction for the different array types is shown and the contributions of the different building blocks are indicated. For reference, the power consumption of the full array size when reducing the number of active co-scheduled UEs is also shown (A 8 UEs, A 4 UEs).

Although with the large array A, the power consumption reduces for eight and four simultaneous UEs (20% for A eight UEs and 30% for A four UEs), the additional power reduction using the smaller arrays K and L is significantly larger - an additional 37% and 64%, respectively. The combination of an adapted array and reduced power consumption during low traffic hours can increase the energy efficiency of a practical system significantly.
To find the resulting energy saving per day, we applied the power consumption of the different array sizes to an ETSI load model [ETSI ES 202706-1 V1.5.1 (2017)], which describes the variation of traffic load over a day. For this example, we associated the traffic profiles of “Busy Load”, “Medium Load” and “Low Load” with the number of 16, eight and four simultaneously active UEs, respectively.

Figure 11. Relative power savings with array adaptation

Figure 12. Relative energy savings with array size adaptation over day

<table>
<thead>
<tr>
<th>Traffic load</th>
<th>Duration hours/day</th>
<th>no. of UEs</th>
<th>Array type</th>
<th>Energy consumption [%]</th>
<th>Array type</th>
<th>Energy consumption [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Busy load</td>
<td>8</td>
<td>16 UEs</td>
<td>A 16UEs</td>
<td>39.6%</td>
<td>A 16UEs</td>
<td>39.6%</td>
</tr>
<tr>
<td>Medium load</td>
<td>10</td>
<td>8 UEs</td>
<td>A 8UEs</td>
<td>39.6%</td>
<td>K 8UEs</td>
<td>24.8%</td>
</tr>
<tr>
<td>Low load</td>
<td>6</td>
<td>4 UEs</td>
<td>A 4UEs</td>
<td>20.9%</td>
<td>L 4UEs</td>
<td>7.4%</td>
</tr>
<tr>
<td>total</td>
<td>24</td>
<td>A</td>
<td>A, K, L</td>
<td>100.0%</td>
<td>A, K, L</td>
<td>71.8%</td>
</tr>
</tbody>
</table>

energy saving per day with array adaptation 28.2%
Figure 12 summarizes the percentage of energy consumption during the hours of different traffic load for the different array size. The energy consumption with array type A and different load conditions over one day is considered as 100%. When using the example array adaptation scheme with array types K and L for medium and low load, respectively, about 30% of the energy can be saved without compromising the performance of the UE.

This deactivation of entire mMIMO transceiver chains is compatible with any other method of the 5G system to save further power, such as micro sleep from using the lean carrier concept and discontinuous transmission (DTX) or disabling carriers completely/reverting to LTE during low loads.

Site Efficiency

Millimeter Wave Massive MIMO on Existing Sites

An important addition to the 3GPP New Radio Air Interface for 5G is the support for carrier frequencies in the mmWave bands, such as 28GHz or 39GHz. The mmWave bands have relatively poor propagation characteristics compared to the traditional sub-6 GHz cellular operating bands. Radio wave propagation in the mmWave bands is much more susceptible to blockage by objects in the environment due to the lack of diffraction as a significant propagation mode. As a result, coverage range and reliability become important factors in judging the effectiveness of a mmWave deployment.

Despite the relatively poor propagation characteristics in the mmWave bands, large user data rates can still be achieved due to the relatively large available bandwidths and through the use of large-scale antenna arrays and Massive MIMO. High gain adaptive beamforming with Massive MIMO can overcome the poor path loss conditions and provide very high data rates for users at favorable locations in the cell.

A key deployment strategy for exploiting the mmWave band is to deploy mmWave access points to add capacity to an existing sub-6GHz deployment, for example, in a city with many users with large data demands. With this strategy, mmWave access points can be deployed on existing sites to provide large data rates to users within a certain distance from the site. Users further away or who have poor mmWave propagation conditions can generally be served with the sub-6 GHz access point. Serving the closer-in users with the mmWave access point can also provide significant benefits to the sub-6 GHz system by freeing up over-the-air resources that would otherwise be assigned to those closer-in users.

Figure 13 shows the downlink performance of a mmWave access point in the urban macro scenario (as defined in 3GPP 38.901) for several inter-site distances (100m, 200m, 500m, 1000m). The system evaluated in Figure 13 was a 512-element access point using hybrid beamforming at 28 GHz with a 200 MHz system bandwidth and a downlink EIRP of 60dBm. The access point array structure was a 16-row, 16-column arrangement of cross-pol antenna elements. The performance of two configurations is shown: one for single-user MIMO and a second for multi-user MIMO. Note that the mmWave system can achieve a reasonably high sector spectral efficiency with inter-site distances up to roughly 200m, after which performance declines and coverage reliability becomes problematic, as indicated by the low cell edge performance. The traffic model in this evaluation was full buffer traffic with an average of 15 continuously active users per site. Note how significant gains can be achieved with MU-MIMO compared to SU-MIMO in these conditions. These results show how mmWave can satisfy the data demands of a dense deployment.
Massive MIMO in Rural Deployments

Another important use case for massive MIMO is in rural deployments where large cell sizes with relatively high tower heights are quite common, although user densities may not be nearly as high as in urban deployments. Massive MIMO has generally been viewed as essential for high user densities and development of the New Radio air interface focused on the urban macro scenario with a relatively small inter-site-distance (200m). As a result, operators in rural areas are asking if Massive MIMO can provide significant performance benefits in rural scenarios where antenna heights are relatively high (e.g., up to 90m), site-to-site distances are rather large (e.g., 8-24km), and user densities are such that the opportunities for simultaneously pairing multiple users (i.e., MU-MIMO) is relatively infrequent.

When considering this question, it is important to understand that the benefits of massive MIMO are two-fold: increasing both capacity and coverage reliability through the use of high gain adaptive beamforming in addition to spatial multiplexing of multiple simultaneous users. The result is that massive MIMO can provide significant performance enhancements even in suburban or rural environments where user densities may be low and/or cell sizes may be relatively large. Even if multi-user operation is relatively infrequent, there can still be significant benefits from the larger arrays due to the higher array gains.

In this section, we present the results of a case study on the performance benefits of Massive MIMO in a rural macro scenario operating at a 2 GHz carrier frequency and a 10 MHz bandwidth. We consider an example with three-sector sites laid out with a site-to-site distance of 8km and 90m base heights. To show the benefits of large-scale arrays, four MIMO configurations are considered in an FDD deployment. The first two configurations are a "legacy" 8-port array having four columns of cross-polarized elements, one configured to perform SU-MIMO with the Rel-10 LTE CSI feedback (denoted as a "basic-LTE" configuration) and the other configured to perform MU-MIMO with the 3GPP Rel-15 NR Type II CSI (denoted as an "optimized NR" configuration).
The second two configurations are “Massive MIMO” configurations each having 32 ports with eight columns of cross-pol elements, one configured to perform MU-MIMO with the 3GPP Rel-15 NR Type I CSI (denoted as a “basic NR” configuration), and the other configured to perform MU-MIMO with the 3GPP Rel-15 NR Type II CSI and advanced regularized zero-forcing precoding (denoted as an “optimized NR” configuration).

In all cases, the transmission scheme is adapted between SU and MU-MIMO transmission with a greedy-pairing algorithm based on proportional fair scheduling metrics. The performance characteristics will be shown in both a full buffer traffic scenario as well as in FTP traffic. In this study, the air interface overhead was the same whether the system was LTE or NR, thereby allowing the results to highlight the MIMO-specific performance aspects. Any overhead advantages provided by the NR air interface would improve the results further compared to an LTE system.

Figure 14 shows the sector spectral efficiency and edge user spectral efficiency in full buffer traffic with an average of 10 active users in each sector of each site in the rural scenario. Figure 15 shows the mean UE spectral efficiency in FTP traffic (FTP Model 1) with three arrival rates. In these figures, we see significant gains from using the 32 port arrays compared to the eight port arrays regardless of the traffic conditions. We also see significant gains from more advanced feedback schemes with advanced precoders, as shown by the differences between the basic configurations and the optimized configurations. Also note that in FTP traffic the relative performance benefits of the 32-port array increase as the FTP packet arrival rate increases.

Figure 14. Sector and edge user spectral efficiency for four MIMO configurations in a Rural Macro scenario in full buffer traffic (30 active UEs per 3-sector site) at 2 GHz. Configurations: (1) 8-port Basic LTE: Rel-10 PMI with SU-MIMO. (2) 8-port Optimized NR: Rel-15 Type II CSI with MU-MIMO and Regularized ZF. (3) 32-port Basic NR: Rel-15 Type I CSI with MU-MIMO and No ZF; (4) 32-port Optimized-NR: Rel-15 Type II CSI, MU-MIMO with Regularized ZF.
In bursty traffic, the average number of users that are co-scheduled with MU-MIMO transmission tends to be relatively low - in the cases shown in Figure 15, the average number of paired UEs ranges between 1.1 and 1.9 where the pairing rate increases with higher arrival rates. In bursty traffic conditions, there tend to be relatively few opportunities to pair users, simply because there is rarely more than one active user at any given time. As arrival rates increase, there are more active users, so MU-MIMO co-scheduling tends to be more frequent at the higher arrival rates.

Also, at the higher arrival rates, we tend to see the less-capable arrays pairing more UEs, on average, than the more-capable arrays, simply because the less-capable arrays cannot serve the active UE population as effectively, which results in more opportunities to pair UEs. By contrast, the average number of co-scheduled UEs in full buffer traffic is significantly higher than in bursty traffic. In the cases shown in Figure 14, the average number of paired users ranges between 3.5 and 5.7, with the more capable arrays tending to pair significantly more UEs than the less capable arrays. With the full buffer traffic model of Figure 14, there are on average 10 active UEs at any given time, which results in much higher pairing rates than in bursty traffic conditions. All this points to the fact that Massive MIMO can provide significant system-level benefits, even in scenarios where the MU co-scheduling capability is rarely needed or used.

Figure 15. Mean user spectral efficiency for four MIMO configurations in a Rural Macro scenario in FTP traffic at 2 GHz

Another important aspect of Massive MIMO is the impact on the over-the-air resource utilization, which is the fraction of the over-the-air resources that are occupied with data transmission. Figure 16 shows the resource utilization for the FTP traffic cases. Resource utilization with full buffer traffic is generally 100% due to the constant presence of 10 active users per sector and, as a result, is not shown in the figures. Note how in Figure 16, resource utilization increases with the arrival rate, but the 8-port arrays have significantly higher resource utilization than the 32-port arrays.
In real deployments, the user experience starts to degrade noticeably as resource utilization gets too high (e.g., over 70%), which may trigger the need to upgrade the deployment to better serve the coverage area – for example either by adding another carrier or by densifying the deployment. If the arrival rate continues to increase, there will be a point where the rate of packets arriving into the system exceeds the spectral efficiency capabilities of the gNB. This will lead to an unstable system with near-zero user throughputs and near-infinite packet delays.

It is important to note that the performance trends shown in this section for a rural macro scenario are very similar to the trends seen in urban macro or urban micro scenarios. With the demand for wireless data constantly increasing over time, larger arrays are more future proof and can significantly delay the time at which the system will need upgrading to deal with the increased data demands. Also, large scale arrays coupled with the New Radio air interface can service much larger cell sizes due to the ability to beamform all over-the-air channels.

Figure 16. Resource Utilization in FTP model 1: four MIMO configurations in a Rural Macro scenario in FTP traffic at 2 GHz
Summary and Outlook

The way to further improve the spectral efficiency of massive MIMO systems is depicted in three innovation categories.

Phase I targets improvements on the (slower timescale) beamforming with an improved grid-of-beams (GoB) concept and especially using Eigenbeamforming, which increases the geometric-mean user rate and in particular boosts the cell edge performance.

Phase II then targets improvements with multiuser MIMO precoding using instantaneous channel knowledge. Algorithms such as block diagonalization (BD) and robust ZF provide roughly a further 30% geometric-mean user rate gain compared to Eigenbeamforming, while the cell edge performance stays the same.

In phase III, distributed MIMO should be introduced, which we have shown produces a further 50% spectral efficiency gains, despite the presence of an already large number of elements in the single-site processing baseline.

Reducing the power consumption of massive MIMO systems is important for OPEX- and CO₂-emissions savings. The Nokia innovation of array configuration adaptation by load-dependent switching-off of transceiver chains can bring down mMIMO power consumption by up to a factor of four in certain scenarios, while keeping the average user rate constant (provided that significantly fewer users are active). For a typical day-night-cycle of varying load, this means a reduction in power consumption of 30%.

Our studies have shown that massive MIMO also provides clear benefits to coverage and throughput in rural deployments, even in situations where there are few opportunities to spatially multiplex multiple users. As seen in both urban and rural deployments, larger arrays are more future proof and can significantly delay the time at which the system will need upgrading to deal with increased data demands. Large scale arrays coupled with the New Radio air interface can service much larger cell sizes due to their ability to beamform all over-the-air channels.

For offloading traffic to higher mmWave bands, the same sites can be shared for both sub-6 GHz as well as e.g. 28 GHz. With 200m inter-site distances, such as those found in cities) more than 10 bps/Hz of traffic can be carried per cell by a mmWave system, freeing up the lower band.

Nokia is a frontrunner in massive MIMO technology, with a 10-year history in active antenna R&D, with first beamforming deployments in 2012. Currently, Nokia has deployed massive MIMO products in LTE and in 5G networks on multiple different frequency bands with major operators globally. Nokia is taking the next steps to further improve massive MIMO capabilities with more powerful ReefShark chipsets and more advanced algorithms using Bell Labs research results.
Appendix

A1: Details on NR type II CSI and ETF

In Rel. 15, NR type II CSI is an eigenvector approximation scheme for CSI feedback, defined up to rank=2 transmission. Assuming $N_1$ is the number of antenna ports in the azimuth direction, $N_2$ is the number of antenna ports in the elevation direction, $L$ is the number of spatial domain (SD) beams per polarization and $N_3$ is the number of subbands. The precoding vector for one layer is written as:

$$ W = W_1 W_2 $$

where $W_1$ is the spatial grid-of-beam matrix of size $2N_1 N_2 \times 2L$. $W_2$ is a matrix of size $2L \times N_3$, it contains the subbands coefficients needed to linearly combine the long-term 2D DFT beams inside $W_1$.

In Rel.16 NR type II CSI, DFT based FD compression scheme is applied on $W_2$, such that the precoding vector can be written as:

$$ W = W_1 \tilde{W}_2 W_f^i $$

where $\tilde{W}_2$ is a matrix of linear combination coefficients (LCC) of size $2L \times M$ and $W_f^i$ is the frequency domain (FD) basis subset matrix of size $M \times N_3$, where $M < N_3$ is the number of FD components per layer. Each of $W_1$ and $W_f^i$ is drawn from a DFT codebook, to achieve the targeted compression in spatial and frequency domain directions respectively, $\tilde{W}_2$ is obtained by applying DFT compression on $W_2$ from Release 15:

$$ \tilde{W}_2 = W_2 W_f $$

Similarly, ETF as discussed in [AVW18, AVW19], exploits the time domain channel sparsity to reduce the UL CSI overhead by applying a projection matrix derived from an IDFT codebook.

Assuming $N_{rx}$ antennas at the UE side and $B$ channel frequency samples, the compressed feedback matrix $G$ of size $2LN_{rx} \times M$, can be obtained from the channel frequency matrix $H$ of size $2LN_{rx} \times B$ and projection matrix $P_s$ of size $B \times M$ as:

$$ G = HP_s $$
Abbreviations

3GPP  3rd Generation Partnership Project
CSI  channel state information
CSI-RS  channel state information reference symbols
CoMP  coordinated multi-point transmission / reception
DFT  discrete Fourier transformation
DL  downlink
D-MIMO  distributed MIMO
EIRP  effective isotropic radiated power
eMBB  enhanced mobile broadband
ETF  explicit time domain feedback
FDD  frequency division duplexing
FR1  frequency range 1 (sub-6GHz)
FR2  frequency range 2
FTP  file transfer protocol
GoB  grid of beams
gNB  gNode B (New radio base station)
IDFT  inverse discrete Fourier transformation
ISD  inter site distance
JT  joint transmission
MIMO  Multiple input multiple output
mmWave  millimeter wave
mMIMO  massive MIMO
MU-MIMO  multiuser MIMO
MRT  maximum ratio transmission
NR  New Radio
PA  power amplifier
PMI  precoding matrix indicator
PRB  physical resource block
RF  radio frequency
RI  rank indicator
Rx  receive
RZF  regularized zero forcing
SINR signal to interference plus noise ratio
SRS sounding reference symbols
SSB synchronization signal block
SU-MIMO single user MIMO
SVD singular value decomposition
TDD time division duplexing
TRP transmission and reception point
Tx transmit
TXRU transmit and receive unit
UE user equipment
UL uplink
UMi urban micro channel
ZF zero forcing

Further Reading


[R1-19] “R1-1902304 “Summary of CSI enhancement for MU-MIMO”, RAN1#96

About Nokia

We create the technology to connect the world. Only Nokia offers a comprehensive portfolio of network equipment, software, services and licensing opportunities across the globe. With our commitment to innovation, driven by the award-winning Nokia Bell Labs, we are a leader in the development and deployment of 5G networks.

Our communications service provider customers support more than 6.4 billion subscriptions with our radio networks, and our enterprise customers have deployed over 1,300 industrial networks worldwide. Adhering to the highest ethical standards, we transform how people live, work and communicate. For our latest updates, please visit us online www.nokia.com and follow us on Twitter @nokia.

Nokia is a registered trademark of Nokia Corporation. Other product and company names mentioned herein may be trademarks or trade names of their respective owners.

© 2021 Nokia
Nokia OYJ
Karakaari 7
02610 Espoo
Finland
Tel. +358 (0) 10 44 88 000

Document code: CID210389