Two-Layer Linear Processing for Massive MIMO on the TitanMIMO Platform

Sébastien Roy
Dept. of Electrical and Computer Engineering
Université de Sherbrooke
e-mail: s.roy17@usherbrooke.ca

January 27, 2015

Massive MIMO is a cornerstone technology in reaching the 5G target of a thousand-fold capacity increase by 2020. The paradigm is based on the fact that if there are enough antennas at the base station (several hundreds to a thousand or more), the so-called *massive* effect is observed whereby simple linear processing (eigenbeamforming and maximal-ratio combining) becomes optimal. This is attractive because such processing is not only extremely simple, it scales linearly with the size of the array and requires very little inter-processor communication. However, while antennas are not in themselves costly, such very large arrays with hundreds or thousands of RF front-ends, A/D and D/A converters are cumbersome and energy-hungry, to say the least. It is therefore of interest to explore the quasi-massive case, where the number of antennas is not sufficient to achieve the massive effect, but still large enough to make full-fledged interference nulling processing — such as minimum mean-square error (MMSE) and zero forcing (ZF) — undesirable because their complexity scales in polynomial fashion, according to the cube of the number of antennas. We show here that applying MMSE or ZF on subsets of antennas and further combining the resulting outputs in a second layer of processing constitutes an attractive approach to achieve good performance with reduced numbers of antennas, while limiting complexity. Furthermore, this approach maps extremely well to the TitanMIMO modular architecture where remote radio head (RRH) units of 8 antennas, each equipped with local baseband processing capability, are aggregated together to form massive MIMO prototyping platforms of various sizes.

1 What is massive MIMO?

The massive MIMO (multi-input, multi-output) paradigm is an evolution of multi-user MIMO (MU-MIMO). It should be understood that MIMO, in current wireless standards such as LTE and 802.11, is typically implemented in its single-user form. That is to say that on a given channel and in a given timeslot, all the base station antennas are used to communicate with a single user terminal, itself being equipped with multiple antennas, where the multiplicity of antennas at both ends allows the creation of multiple data streams in space, thus multiplying the link capacity by a significant factor. As it is easier to have a lot of antennas (for size and cost reasons) at the base station compared to the handset, these additional degrees of freedom can be used to communicate with multiple users at the same time, giving rise to MU-MIMO. However, this is a much more difficult problem given that the multiple users addressed simultaneously cannot easily perform joint processing in order to eliminate the inter-user interference created with this method.

Therefore, while MU-MIMO is supported in LTE release 8, the precoding scheme therein does not completely address the interference problem. It follows that efficient implementation requires clever processing beyond what is dictated by the standard, and many system designers are skeptical with this route. The reason is that it is much simpler to maximize single-user MIMO (SU-MIMO) throughput, thus liberating the channels sooner for other users, rather than attempt MU-MIMO while the system gains are similar.

Given this state of affairs, neither SU-MIMO nor MU-MIMO is sufficiently powerful to achieve the 1000× capacity increase demanded by 5G. The seminal paper by Marzetta [1] introduced the concept of "massive MIMO" in 2010 (also referred to as large-scale antenna systems or LSAS), generating immediate interest and numerous other papers [2]–[4]. It constitutes a theoretical and asymptotic analysis of a multi-cell scenario where a population of single-antenna terminals are served by cellular base stations having an infinite number of antennas. While some real-world constraints are not considered, this work provides useful insights into the benefits and drawbacks of LSAS. Namely, when the number of base station antennas is allowed to tend towards infinity,

1. the effect of uncorrelated noise and fast fading vanish;
2. throughput and the number of terminals become independent from the size of the cells;
3. the required transmitted energy tends towards zero (due to infinite array gain);
4. multi-user interference vanishes; and
5. very simple forms of detection and precoding, namely matched filtering and eigenbeamforming, become optimal.

However, such theoretical fundamental benefits cannot be achieved without overcoming multiple practical hurdles. Known issues affecting massive MIMO include

1. **Pilot contamination**: In massive MIMO, it is unrealistic to operate in FDD mode and to use some sort of sounding / feedback technique to obtain downlink channel estimates, given

the staggering overhead this would entail for such a large array. TDD operation is generally assumed, with a frame size such that decent downlink channel estimates can be obtained by reciprocity from the uplink channel estimates. Typically, the latter are obtained during a dedicated training interval, or, equivalently using pilot symbols. The pilots or training sequences can be made orthogonal or quasi-orthogonal among users for a single cell, but will necessarily be contaminated by transmissions (re-use of training sequences) from surrounding cells. This effect, which does not diminish with the size of the array L, is widely recognized as one of the main practical capacity limitations of the massive MIMO paradigm.

2. **Array scale**: The sheer array scale required to achieve the true massive effect, whereby eigenbeamforming and maximal-ratio combining can be leveraged, involves staggering size, cost, and energy consumption considerations.

3. **Array coherence**: Appropriate synchronization, calibration, and phase alignment accross a large-scale array, of hundreds of antennas or more, pose serious practical challenges.

To summarize, the true massive effect implies that eigenbeamforming and maximal-ratio combining are optimal, transmit power can be made arbitrarily small, white noise, channel estimation error and interference vanish, and capacity is limited by pilot contamination. This is known to result when the base station array size L tends towards ∞ while the number of addressed user antennas M remains fixed. Obviously, a limit must be imposed on the size of the array in practice, as expressed by hurdle 2 above. Hoydis, ten Brink and Debbah have asked the important question "How many antennas to we need?" in practice in order to achieve a significant percentage of the ideal massive MIMO performance [5]. They show therein that in certain scenarios, ZF and MMSE processing can attain the performance level of eigenbeamforming / MRC with an order of magnitude fewer antennas. However, the resulting array must still be large with respect to the user population of size M, and the complexity of ZF and MMSE raises serious concerns given that it scales according to L^3 and that processing accross the entire array becomes tightly intercoupled, requiring rapid and flexible connections between local baseband processors.

A key paper [6] describes Argos, the first (and one of the few) hardware implementation of an LSAS hardware system in the spirit of massive MIMO. This reference is crucial because in the process of making a working system, based on a modular, scalable architecture, the authors had to address (and describe therein) the main practical issues associated with such an undertaking. The system needed to be scalable both in terms of numerical complexity and in terms of data routing requirements. Indeed, as the array becomes very large, it becomes impractical (in terms of hardware complexity and / or delay) to gather data from all antennas in one central location for weight calculation, detection and / or channel estimation. Thus, all these tasks must ideally be performed locally at each antenna, and their numerical complexity should scale linearly with the number of antennas. The prototype system comprises one FPGA processing board for every 4 antennas, so its computation power does scale linearly as well. In order to achieve this, the authors used conjugate (matched) beamforming with local power scaling, so that channel estimation and weight computation can be performed locally. Furthermore, internal calibration is implemented relative to antenna 1, and this is shown to be adequate to compensate for the effect of RF component imperfections on channel reciprocity. However, their results show that zero-forcing beamforming yields a capacity advantage by a factor of 2 to 4 for the relatively small number of antennas (16 to 64) used in the experiment. This confirms the results of [5] discussed above. Thus, many more antennas (100s) would be required for conjugate beamforming to become optimal, yet zero-forcing has a numerical complexity that scales with L^3 (due to matrix inversion) and requires data routing to a central controller for weight computation.

The Nutaq TitanMIMO platform [7] represents a significant evolution with respect to Argos in many respects. It provides one large Virtex-6 FPGA for every 8 antennas, with all modules being linked in a user-defined topology with high speed point-to-point links (up to 7x 20Gbps links per module). The latter enable the testbed to be used in a wide range of scenarios making it a perfect fit for massive MIMO/multi-layer processing research-oriented projects. Figure 1 shows an example of a TitanMIMO system which supports up to 64 RF transceivers, using an RRH-only topology.

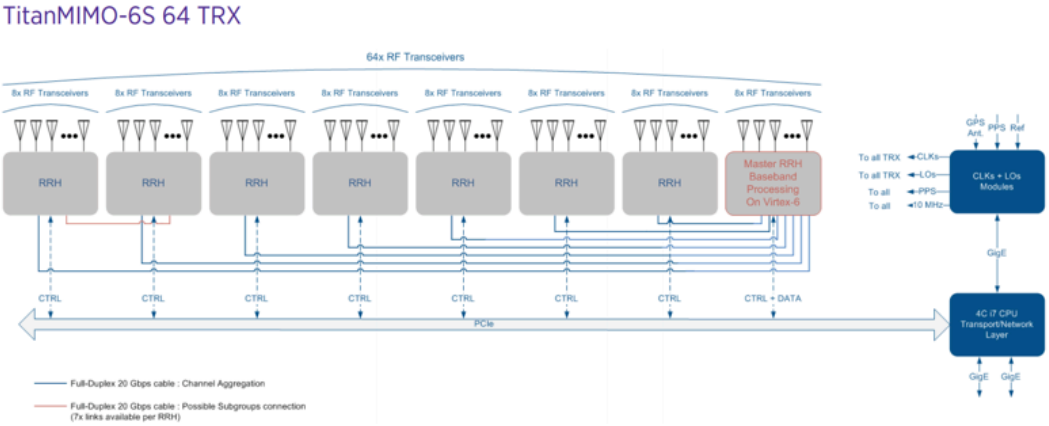
TitanMIMO-6S 64 TRX
64x RF Transceivers
8x RF Transceivers
RRH
Master RRH Baseband Processing On Virtex-6
GPS Ant.
PPS
Ref
To all TRX
CLKs
To all TRX
LOs
To all
PPS
To all
10 MHz
CLKs + LOs Modules
GigE
CTRL
CTRL + DATA
PCIe
4C i7 CPU Transport/Network Layer
GigE
GigE
Full-Duplex 20 Gbps cable : Channel Aggregation
Full-Duplex 20 Gbps cable : Possible Subgroups connection (7x links available per RRH)

Figure 1: TitanMIMO-6S 64 TRX configuration

Since each module (RRH) possesses 7 high speed point-to-point links, a large number of different topologies can be implemented. However, as the number of antennas increases, additional processing power is required and there comes a point where the local processing of the RRH modules may not suffice. However, the scalability of the TitanMIMO can be pushed further by adding one or more octal-FPGA central processing boards (Kermode XV6) as shown in Figs. 2 and 3.

This scalability can attain up to 1000 TX/RX channels and the additional processing power brought by the Octal-FPGA processing boards enables the implementation of higher complexity algorithms such as ZF and MMSE.

2 Two-layer processing

Given large-scale arrays of manageable sizes in practice, the above discussion highlights the need for algorithms that are simpler and that scale better than full-fledged ZF or MMSE, while offering

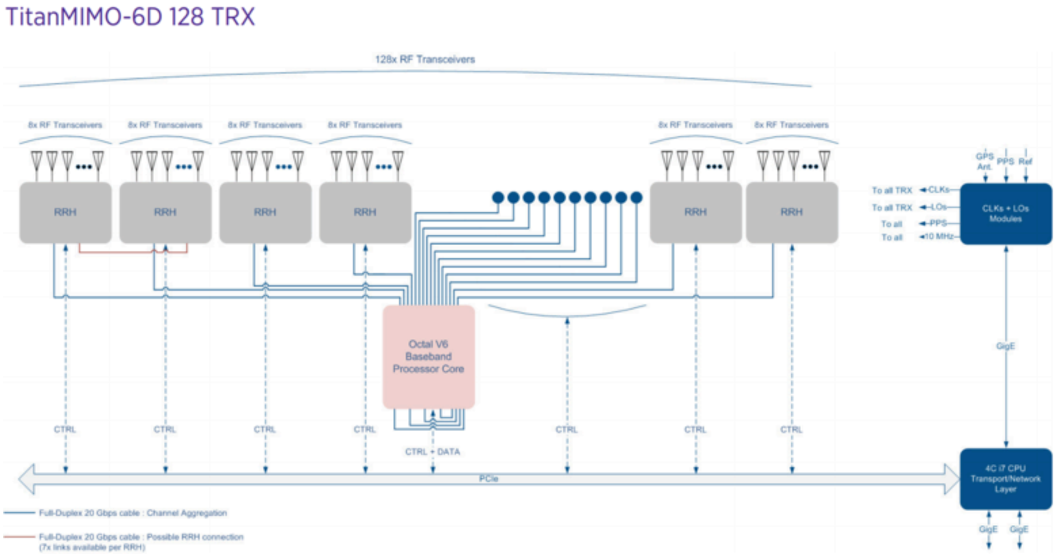
TitanMIMO-6D 128 TRX
128x RF Transceivers
8x RF Transceivers
RRH
Octal V6 Baseband Processor Core
CTRL
CTRL + DATA
PCIe
GPS Ant.
PPS
Ref
To all TRX
CLKs
To all TRX
LOs
To all
PPS
To all
10 MHz
CLKs + LOs Modules
GigE
4C i7 CPU Transport/Network Layer
Full-Duplex 20 Gbps cable : Channel Aggregation
Full-Duplex 20 Gbps cable : Possible RRH connection (7x links available per RRH)

Figure 2: TitanMIMO-6D 128 TRX configuration

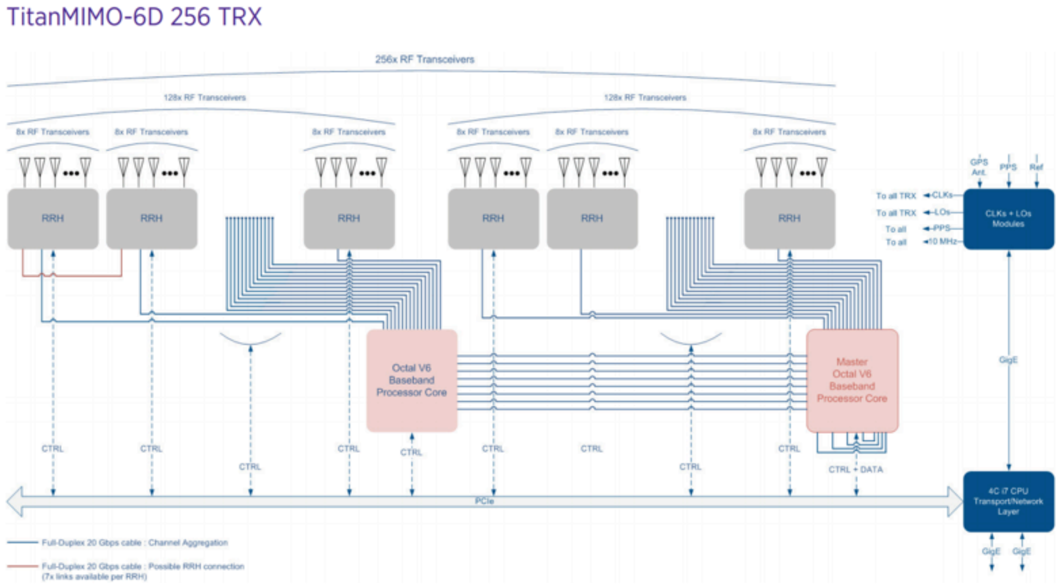
TitanMIMO-6D 256 TRX
256x RF Transceivers
128x RF Transceivers
8x RF Transceivers
RRH
Octal V6 Baseband Processor Core
Master Octal V6 Baseband Processor Core
CTRL
CTRL + DATA
PCIe
GPS Ant.
PPS
Ref
To all TRX
CLKs
To all TRX
LOs
To all
PPS
To all
10 MHz
CLKs + LOs Modules
GigE
4C i7 CPU Transport/Network Layer
Full-Duplex 20 Gbps cable : Channel Aggregation
Full-Duplex 20 Gbps cable : Possible RRH connection (7x links available per RRH)

Figure 3: TitanMIMO-6D 256 TRX configuration

better interference nulling than eigenbeamforming / MRC. One promising approach on the reception side is to form subsets of antennas of a given size N, to apply ZF or MMSE at the subset level (the first processing layer), and to subsequently combine the resulting outputs using MRC (second processing layer). In this approach complexity is essentially proportional to N^3, instead of

L^3, where N can be made much smaller than L, according to the desired complexity-performance tradeoff. It should be noted that while N is assumed fixed in the following, it is entirely possible to have subsets of unequal size, which might in fact be necessary for certain array sizes L depending on the number of desired subsets.

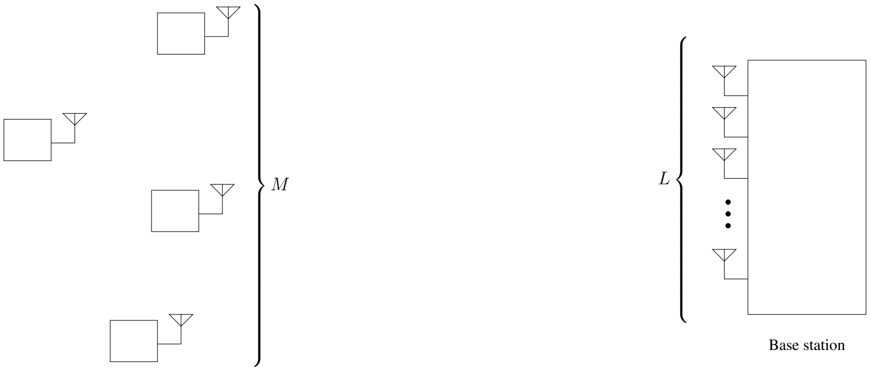

Figure 4: Typical massive MIMO scenario: the base station has L antennas and is servicing M single-antenna terminals, where $L \gg M$.

Figure 4 shows a typical massive MIMO scenario where the base station is equipped with a large number L of antennas and is servicing a population of M single-antenna user terminals on any given channel, where L is assumed to be at least an order of magnitude greater than M. Thus, the complex baseband model of this MIMO link is given by

$$\mathbf{x} = \mathbf{H}\mathbf{s} + \mathbf{n}, \tag{1}$$

where $\mathbf{x} \in \mathbb{C}^{L\times 1}$ is the received signal and $\mathbf{H} \in \mathbb{C}^{L\times M}$ is the Rayleigh-fading channel matrix comprised of independent, identically distributed (i.i.d.) zero-mean circularly symmetric complex Gaussian coefficients denoted by $h_{ml} \sim N(0,1)$ for $1 \le m \le M$, $1 \le l \le L$. It is assumed that the M data streams have equal power. In other words, $\mathbf{s} \in \mathbb{C}^{M\times 1}$ is characterized by the following covariance matrix:

$$\mathbf{R}_{\mathbf{xx}} = \left\langle \mathbf{x}\mathbf{x}^H \right\rangle = P\mathbf{I}_M, \tag{2}$$

where $\langle\cdot\rangle$ is the expectation operator, $(\cdot)^H = (\cdot)^{*T}$ is the Hermitian transpose (or conjugate transpose), and $\mathbf{I}_M$ is the $M \times M$ identity matrix. Finally, the thermal white Gaussian noise vector $\mathbf{n} \in \mathbb{C}^{L\times 1}$ is also circularly symmetric such that $\mathbf{n} \sim N\left(0, \sigma_n^2\mathbf{I}_L\right)$.

For conventional linear processing on the uplink, the vector of estimates of the transmitted signals prior to detection is given by

$$\mathbf{z} = \mathbf{W}\mathbf{x}, \tag{3}$$

where each entry of $\mathbf{z} \in \mathbb{C}^{M\times 1}$ corresponds to the same entry of $\mathbf{s}$, and $\mathbf{W} \in \mathbb{C}^{M\times L}$ is the weight combining matrix. It should be noted that the mth row of $\mathbf{W}$ is the weight combining vector for user m's signal.

If we assume perfect channel knowledge, the weight combining matrix $\mathbf{W}$ for MRC, ZF and MMSE is given by

$$\begin{aligned} \mathbf{W}_{MRC} &= \mathbf{H}^H, &(4)\\ \mathbf{W}_{ZF} &= \left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{H}^H, &(5)\\ \mathbf{W}_{MMSE} &= \left(\mathbf{H}^H\mathbf{H} + \frac{\sigma_n^2}{P}\mathbf{I}\right)^{-1}\mathbf{H}^H. &(6) \end{aligned}$$

It can be seen that both ZF and MMSE require matrix inversion, which for large array size L becomes problematic given that the complexity is $O(L^3)$. Meanwhile, MRC is much simpler to implement but requires a much larger array to yield acceptable performance.

Figure 5 shows the proposed 2-layer processing scheme for reception (uplink) at the base station. Here, the processing for a single user is illustrated and the output is one of the entries of vector $\mathbf{z}$. It should be understood that a similar structure can be leveraged for transmission (downlink) with identical benefits. Therefore, we focus herein and without loss of generality on the uplink, the extension to downlink processing being straightforward.

With this architecture, all antennas in the array are equipped with an RF front-end (including A/D and D/A conversion) and first layer processing in each group is performed in the digital domain. The first layer subset processors implement either ZF or MMSE combining in order to reduce inter-user interference. Thus, we have, for all users

$$\mathbf{y}_k = \mathbf{W}_k\mathbf{x}_k, \tag{7}$$

where $1 \leq k \leq K$, $\mathbf{x}_k$ is the portion of the received signal vector $\mathbf{x}$ corresponding to the kth subset, and $\mathbf{W}_k$ is the weight combining matrix associated with the latter. Thus, $\mathbf{W}_k$ is computed in one of two ways, depending on whether ZF or MMSE processing is desired, i.e.

$$\mathbf{W}_k = \begin{cases} \mathbf{W}_{k,ZF} = \left(\mathbf{H}_k^H\mathbf{H}_k\right)^{-1}\mathbf{H}_k^H & \text{for ZF,}\\ \mathbf{W}_{k,MMSE} = \left(\mathbf{H}_k^H\mathbf{H}_k + \frac{\sigma_n^2}{P}\mathbf{I}\right)^{-1}\mathbf{H}_k^H & \text{for MMSE,} \end{cases} \tag{8}$$

where $\mathbf{H}_k \in \mathbb{C}^{N\times M}$ is the portion of the channel matrix $\mathbf{H}$ corresponding to the kth subset.

To apply MRC in the second layer, we recall that the MRC concept is akin to a spatial matched filter, i.e. each input is co-phased and weighted proportionally to its SINR. It should be noted, however, that all inputs are already co-phased as a result of first-layer processing.

For a given user m, we have

$$\begin{aligned} y_{mk} &= \mathbf{w}_{mk}^H\mathbf{x}_k, \\ &= \mathbf{w}_{mk}^H\mathbf{H}_k\mathbf{s} + \mathbf{w}_{mk}^H\mathbf{n}_k, &(9) \end{aligned}$$

where the mth row of $\mathbf{H}_k$ will henceforth be denoted $\mathbf{h}_{mk}$.

The final output of the two-layer processor for user m is given by

$$z_m = \bar{\mathbf{w}}_m^H\mathbf{y}. \tag{10}$$

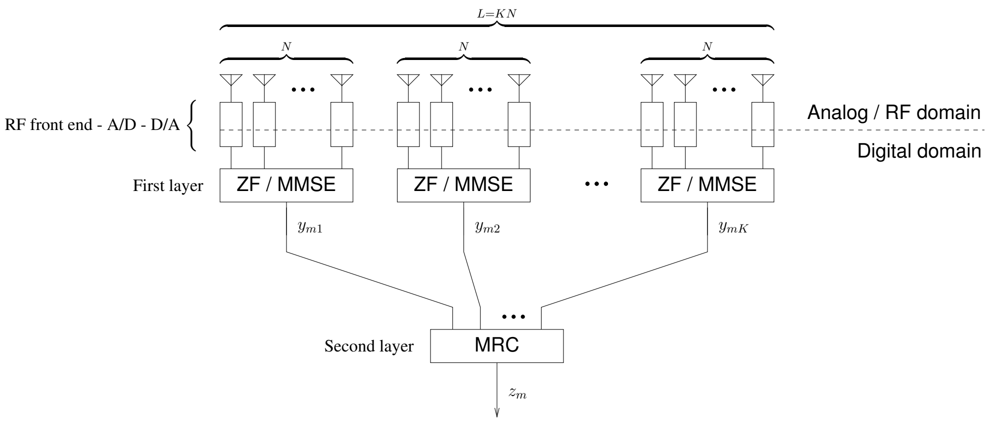

Figure 5: Two-layer processing scheme for the uplink at the base station of a massive MIMO system

If zero-forcing is used in the first layer, we have

$$\bar{\mathbf{w}}_m = \begin{bmatrix} \sqrt{\mathbf{h}_{m1}^H \left(\mathbf{H}_1^H \mathbf{H}_1\right)^{-1} \mathbf{h}_{m1}} \\ \sqrt{\mathbf{h}_{m2}^H \left(\mathbf{H}_2^H \mathbf{H}_2\right)^{-1} \mathbf{h}_{m2}} \\ \vdots \\ \sqrt{\mathbf{h}_{mK}^H \left(\mathbf{H}_K^H \mathbf{H}_K\right)^{-1} \mathbf{h}_{mK}} \end{bmatrix}, \tag{11}$$

where $\mathbf{h}_{mk}$ is the mth column of $\mathbf{H}_k$ or, equivalently, the channel coefficient vector for user m on subset k.

Similarly, if MMSE is used in the first layer, we have

$$\bar{\mathbf{w}}_m = \begin{bmatrix} \sqrt{\mathbf{h}_{m1}^H \left(\mathbf{H}_1^H \mathbf{H}_1 + \frac{\sigma_n^2}{P}\mathbf{I}\right)^{-1} \mathbf{h}_{m1}} \\ \sqrt{\mathbf{h}_{m2}^H \left(\mathbf{H}_2^H \mathbf{H}_2 + \frac{\sigma_n^2}{P}\mathbf{I}\right)^{-1} \mathbf{h}_{m2}} \\ \vdots \\ \sqrt{\mathbf{h}_{mK}^H \left(\mathbf{H}_K^H \mathbf{H}_K + \frac{\sigma_n^2}{P}\mathbf{I}\right)^{-1} \mathbf{h}_{mK}} \end{bmatrix}. \tag{12}$$

If the signal-to-interference-plus-noise ratio (SINR) at the output of subset k for user m is denoted γ_{km}, then the final combined output at the second layer for user m will have an SINR of

$$\gamma_m = \sum_{k=1}^{K} \gamma_{km}. \tag{13}$$

3 Performance comparisons

It should be noted that while many papers on massive MIMO present asymptotic, i.e. approximate, results in part due to the fact that simulations can become very time consuming and cumbersome at large array sizes, all performance results herein for MRC and ZF combiners are exact.

It has been established that as the number of antennas grows to infinity, thanks to the "massive effect," MRC becomes the optimal combining method. However, at more modest array sizes, MRC suffers from an error floor due to the fact that it does not null interference completely, regardless of SNR. This is why ZF and MMSE combiners, which specifically target interferers, can achieve the same performance as MRC at a much smaller array size.

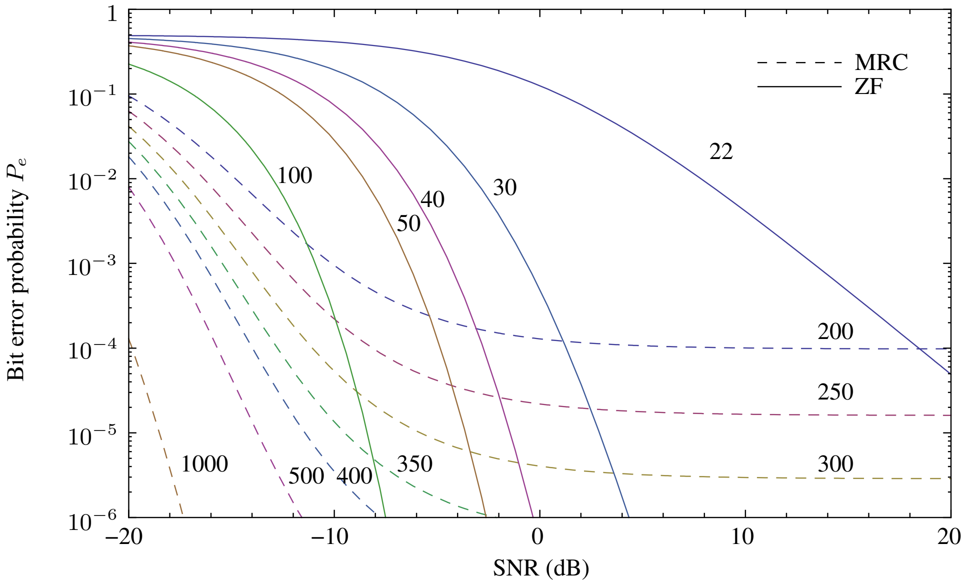

Figure 6: Bit error probability versus SNR for MRC and ZF for various array sizes L when the number of received signals $M = 20$.

Figure 6 compares the error performance of MRC and ZF combiners for various array sizes while assuming a number of transmit antennas of $M = 20$. For analytical convenience (and without loss of generality since the trends and relationships between curves would be similar for any modulation type), DPSK modulation is assumed. It can be observed that to reach a performance level of $P_e = 10^{-3}$ at a SNR of -10 dB (or $P_e = 10^{-4}$ at a SNR of 10 dB), at least 200 antennas are required for MRC. To attain $P_e = 10^{-3}$ at a SNR of -10 dB with ZF, however, 100 antennas is more than enough. And, because there is no error floor, between 22 and 30 antennas would be adequate to reach $P_e = 10^{-4}$ at 10 dB. Observing the crossing points of the curves is most informative. For example, at a SNR of 1.5 dB, a 30-antenna ZF array yields the same error probability as a 200-antenna MRC one. That is nearly a factor of 7 in array size. Likewise, at -3 dB, both a 50-antenna ZF array and a 350-antenna MRC array yield an error probability of 10^{-6}. Here, we have exactly a factor of 7.

It should be noted that the error performance of ZF and MMSE behaves in a very similar way, i.e. the slope of their bit error rate (BER) curves is the same. However, MMSE has a slight advantage, typically on the order of a few dBs, which translates to a shift of the BER curve to the left. It was shown in [8] that there exists a performance gain in MMSE with respect to ZF that does not diminish with increasing SNR. However, it is known that training converges faster with ZF than with MMSE, thus implying that ZF may not suffer as much in the presence of channel estimation errors.

Figure 7 compares the performance of various types of combiners, including 3 instances of two-layer architectures with a number of signals $M = 16$. Given the TitanMIMO RRH size of 8, group sizes of 16 were chosen for the two-layer schemes. Thus, each subset processor in layer 1 can be implemented accross a pair of RRH units. It can be observed that to reach $P_e = 10^{-3}$ at an SNR of 0 dB or better, at least 128 antennas are required with MRC. This same level of performance is attained with only 32 antennas divided into 2 sets of 16 with ZF / MRC 2-layer processing at an SNR of 15 dB. If instead MMSE / MRC processing is used with 2 sets of 16, an SNR of 6 dB is sufficient. It can be seen that the 4×16 ZF / MRC curve has a steeper curve than the 2×16 schemes, but it is poorer than the MMSE / MRC 2×16 processor below 9 dB.

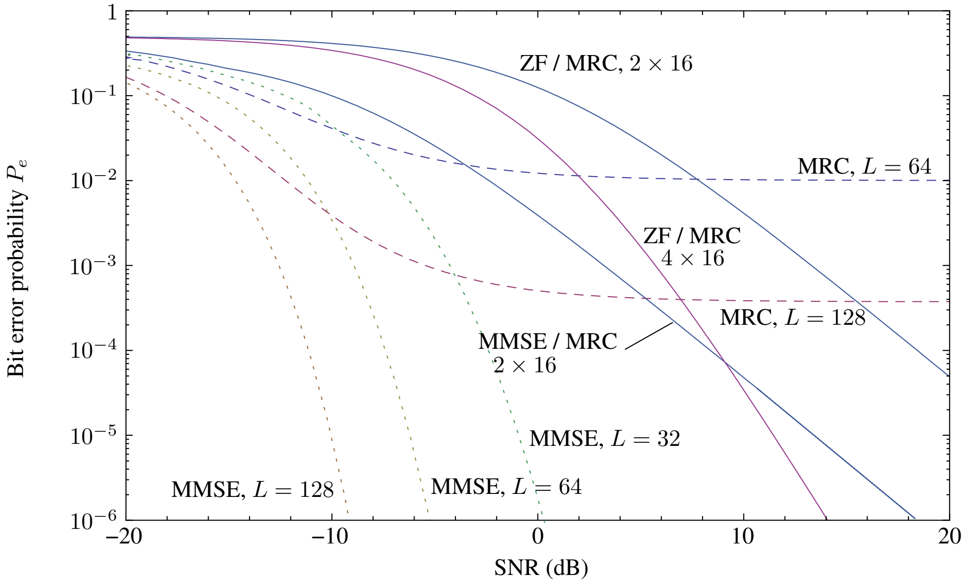

Figure 7: Bit error probability versus SNR for various combiners when the number of received signals $M = 16$.

Figure 8 illustrates the performance of a set of combiners, where the 2-layer schemes are based on a subset size of $N = 8$, with a number of signals $M = 8$. In this case, 64 antennas is enough for MRC to reach a performance level of 10^{-3} with a SNR greater than 0 dB. For the two layer schemes, there is a clear difference in performance and curve slope between 1, 2 and 4 groups of 8.

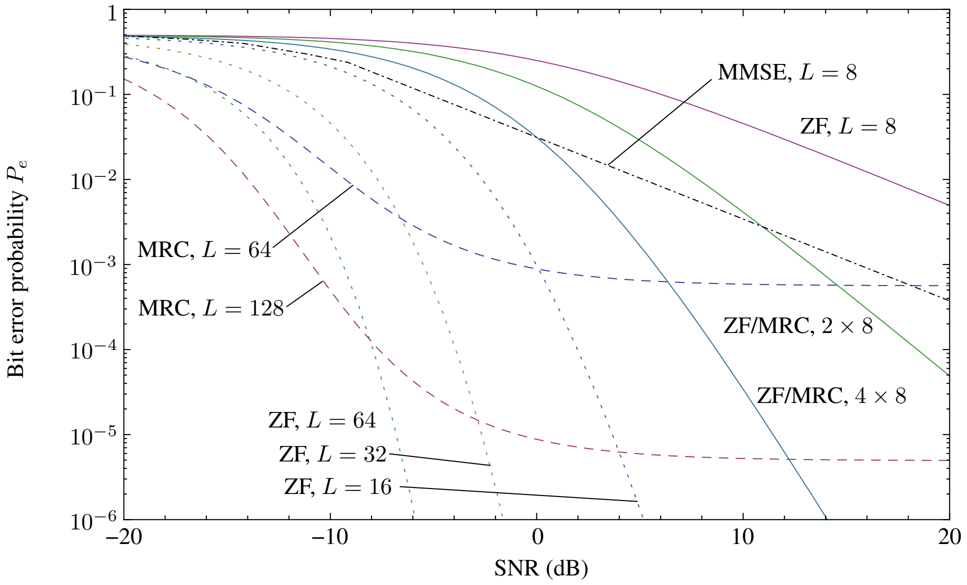

Figure 8: Bit error probability versus SNR for various combiners when the number of received signals $M = 8$.

Figure 9 shows a different angle to performance comparison. Here, the number of impinging signals is $M = 22$ and the subset size is either $N = 24$ (equivalent to 3 RRH modules in a TitanMIMO system) or a multiple thereof. While maintaining the array size at $L = 192$, we vary the number of subsets from 1 to 8. Curves for MRC with $L = 200$ and ZF with $L = 24$ are provided for reference purposes. Going from full ZF with 192 antennas (with a complexity of $O(192^3)$) to 2 subsets of 96 antennas (with a complexity of $O(96^3)$), we see that the performance penalty is less than 1 dB while the complexity is reduced by a factor of $2^3/2 = 4$. Likewise, with 4 subsets of 48, the performance penalty with respect to full ZF is approximately 2.5 dB while the complexity is reduced by a factor $4^3/4 = 16$. With 8 subsets of 24, the performance penalty grows to 8 dB while the complexity reduction is on the order of $8^3/8 = 64$.

4 Prototyping with TitanMIMO platform

The TitanMIMO architecture is designed to be modular and scalable, according to requirements. Processing is naturally distributed since each remote radio head (RRH) module of 8 RF transceivers is equipped with its own Virtex-6 FPGA (Perseus 611x card), and the collection of RRH modules can be linked to one or more Kermode XV6 modules, each being equipped with superlative computing power (8 large FPGAs from the Virtex-6 family) and thus play the role of central processin unit(s).

As with all parallel computing systems, the performance keystone resides in the bandwidth and flexibility of the interconnection network. This is precisely one of the great strenghts of the Ti-

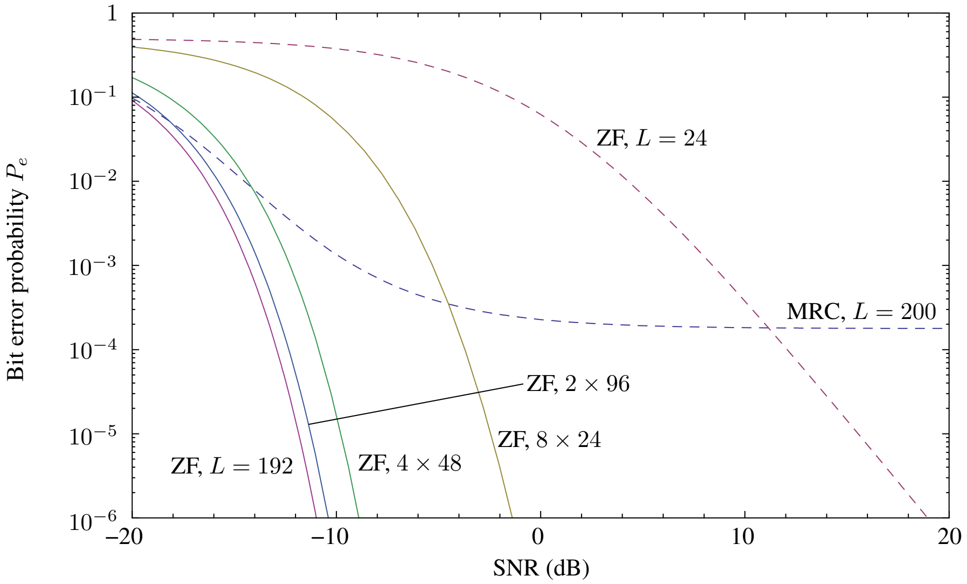

Figure 9: Bit error probability versus SNR for an array of $L = 192$ antennas with 1, 2, 4 and 8 subsets when $M = 22$.

tanMIMO architecture. Indeed, each RRH module has 7 high-speed peer-to-peer connections of 20 Gps, while the Kermode XV6 boards have 32. These connections can be linked point-to-point according to any user-designed topology, thus providing a maximum of flexibility to adapt to a given application.

Thus, the TitanMIMO architecture can be summarized according to 3 main elements:

1. local processing resources (RRH modules and associated FPGAs);

2. global processing resources (Kermode boards);

3. abundant and flexible interconnect resources.

While such an architecture is flexible enough to support very demanding MIMO schemes, it is especially well-suited to two-layer architectures as described here (as well as other variants in [9] [10] [11]). The two-layer concept is designed to result in simpler practical implementations and to enable natural scaling to larger array sizes. This resolves a major issue with classical linear processing algorithms which either a- require too many antennas (MRC) or b- have complexities that explode with increasing array size (ZF and MMSE).

Since the TitanMIMO architecture provides local processing resources via the RRH modules, the subset processing aspect of the two-layer approach is a natural fit. The optimal and most natural solution in terms of implementation complexity is to use subset sizes of 2, 4 or 8 antennas, since an integer number of subsets then fits into each Perseus card. In smaller TitanMIMO systems, it is also feasible to implement two-layer schemes without a Kermode board, using one of the RRH modules as the second layer baseband processor (see Figure 10).

In the case of larger systems, the second-layer processing can be moved to one or more Kermode boards. For example, Figure 11 shows a configuration where the first layer processing is performed on subsets of size $N = 8$, with one subset being mapped to each RRH. However, by exploiting the availability of abundant interconnect resources, it is also feasible to split subset processing accross two or more RRH modules. Thus, Figure 12 illustrates a configuration comprised of 4 subsets of 10 antennas and 2 subsets of 9 antennas; every subset is here split among 2 RRH modules and each RRH module is involved in 2 subsets. With direct connections between neighbors, two approaches are possible: 1- use the rapid links to concentrate all the data within a single RRH for each subset, which then becomes the local baseband processor for the said subset; 2- somehow split the processing tasks for a given subset across two RRH modules (and therefore 2 FPGAs) while setting up the appropriate message passing structure between them.

It is noteworthy that when the second layer is implemented within a Kermode board, much more sophisticated processing than MRC can easily be implemented, including maximum-likelihood, given that the total number of subsets is much smaller than the array size.

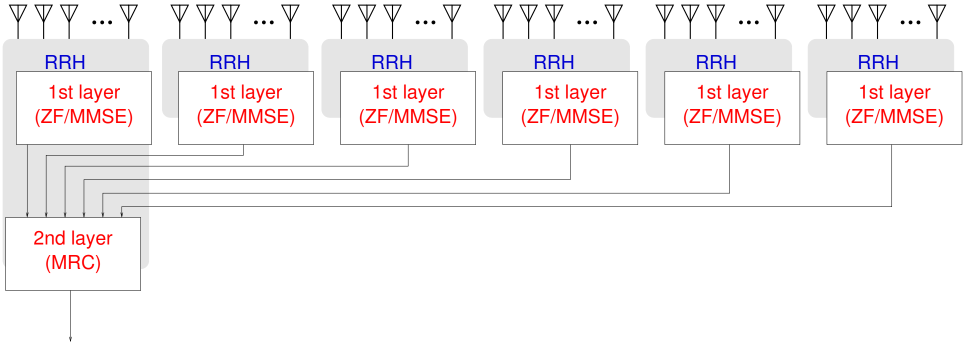

Figure 10: Two-layer processing with subsets of size 8 on a 48-antenna TitanMIMO system without a Kermode XV6 board

5 Conclusion

Two-layer receive processing, whereby a first processing layer consists of splitting the array into a number of subsets of manageable size and performing either ZF or MMSE processing on each subset, and a second processing layer consists of combining the outputs of all subsets according to their respective SNR (MRC), is shown to be a promising avenue for practical implementation of massive MIMO systems. Classical combining approaches, namely MRC, ZF, and MMSE, either a- require too many antennas (MRC) or b- scale very poorly with increasing array size (MMSE, ZF). The two-layer approach fills the performance gap and provides most of the performance advantage of ZF and MMSE while preserving the complexity scalability of MRC. This enables a drastic re-

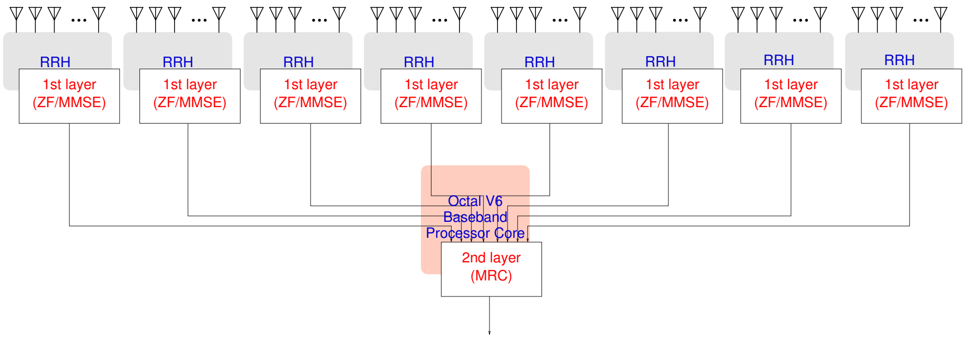

Figure 11: Two-layer processing with subsets of size $N = 8$ on a 64-antenna TitanMIMO system

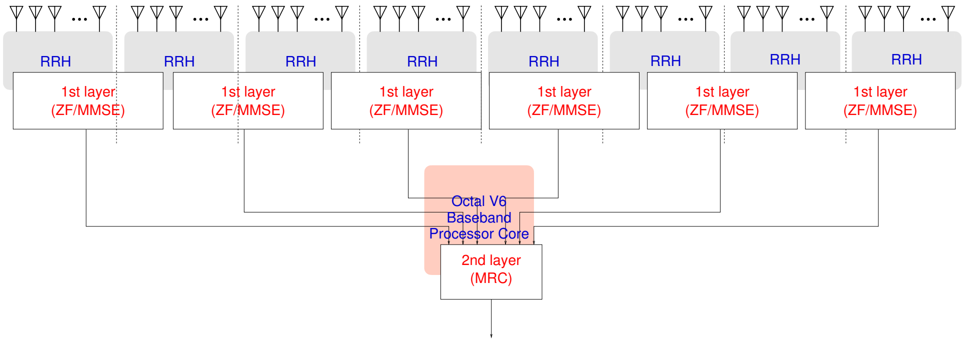

Figure 12: Two-layer processing with subsets of size 8 and 9 on a 64-antenna TitanMIMO system

duction in the number of antennas and RF chains with respect to MRC for similar performance levels. Since this approach is inherently modular, it maps naturally to the TitanMIMO architecture. Furthermore, a TitanMIMO with two-layer processing can be scaled up with predictable and manageable complexity growth. Thus, 2-layer processing can be implemented on a small system which can be incrementally upgraded to larger array sizes with minimal changes to FPGA configurations.

References

[1] T. L. Marzetta, "Noncooperative cellular wireless with unlimited num- bers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[2] A. Pitarokoilis, S. K. Mohammed, and E. G. Larsson, "On the Optimality of Single-Carrier Transmission in Large-Scale Antenna Systems," *IEEE Wireless Commun. Lett.*, vol. 1, no. 4, pp. 276-279, Aug. 2012.

[3] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson,

"Scaling up MIMO: Opportunities and Challenges with Very Large Arrays," *IEEE Signal Process. Mag.*, to appear, 2012.

[4] B. Gopalakrishnan and N. Jindal, "An Analysis of Pilot Contamination on Multi-User MIMO Cellular Systems with Many Antennas," *Proc. Signal Processing Advances in Wireless Communications (SPAWC)*, San Francisco, CA, June 2011.

[5] J. Hoydis, S. Ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[6] C. Shepard, H. Yu, N. Anand, L. E. Li, T. L. Marzetta, R. Yang, and L. Zhong, "Argos: practical many-antenna base stations," in *Proc. ACM Int. Conf. Mobile Computing and Networking (MobiCom)*, Aug. 2012.

[7] M. Brown and M. Turgeon, "TitanMIMO: A 100x100 Massive MIMO Testbed Based on xTCA Standards," white paper, Nutaq Inc., www.nutaq.com.

[8] Y. Jiang, M. K. Varanasi, and J. Li, "Performance analysis of ZF and MMSE equalizers for MIMO systems: an in-depth study of the high SNR regime," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2008–2026, April 2011.

[9] S. Roy, "Dual layer switch-and-examine / selection diversity receivers," *IEEE Vehicular Technology Conference (VTC) – Spring*, Dresden, Germany, June 2–5, 2013.

[10] S. Roy, "Performance analysis of hierarchical selection diversity combining in Rayleigh fading," *International Conference on Networking and Communications (ICNC)*, San Diego, U.S.A., Jan 28–31, 2013.

[11] S. Roy, "Hierarchical Selection Diversity Combining Architecture," *IEEE Int. Conf. on Commun. (ICC)*, Ottawa, Canada, June 10–15, 2012.

Biography

Sébastien Roy is professor at the Departement of Electrical and Computer Engineering at Université de Sherbrooke, Sherbrooke, Canada, as well as founder and head of the World-Machine Interface Research Group (WMIRG). He has published over 120 journal and conference papers, most of which are in the field of wireless communications and, more specifically, antenna arrays and MIMO systems. Active in technology transfer and industrial consulting, he holds 7 patents in the US and worldwide. He recently received the *VTS Conference Chair Award*, jointly with co-chair André Morin, for chairing the Vehicular Technology Conference (VTC) Fall 2012 in Quebec City.