# Hardware Accelerators for Massive MIMO

**Hemanth Prabhu, Lund University, LUND, Sweden**

# Outline

- Why hardware accelerator
- Detector in LUMAMI
- ASIC implementation of pre-coders
- Pre-coding strategies for hardware imparity
- Conclusion

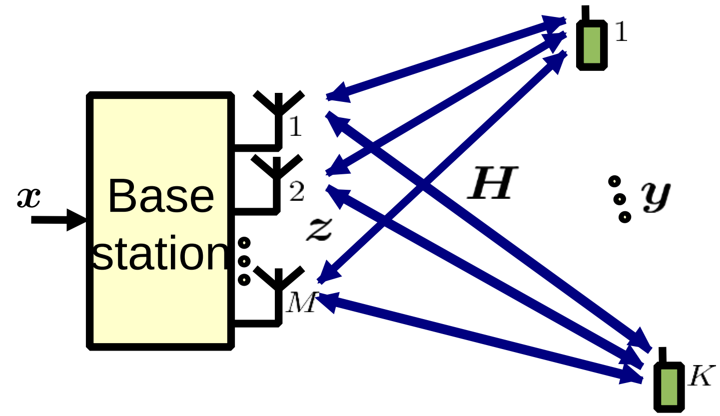
# Linear pre-coders and detectors

- Linear Pre-coder

Maximum-ratio transmission (MRT)

$$z = H^H x$$

Hermitian transpose of channel




Zero-forcing (ZF)

$$z = H^+ x$$

Pseudo-inverse of channel

- Similarly uplink can be performed using linear detectors.

# Processing Cost

- Complexity break-down for ZF detector/pre-coder

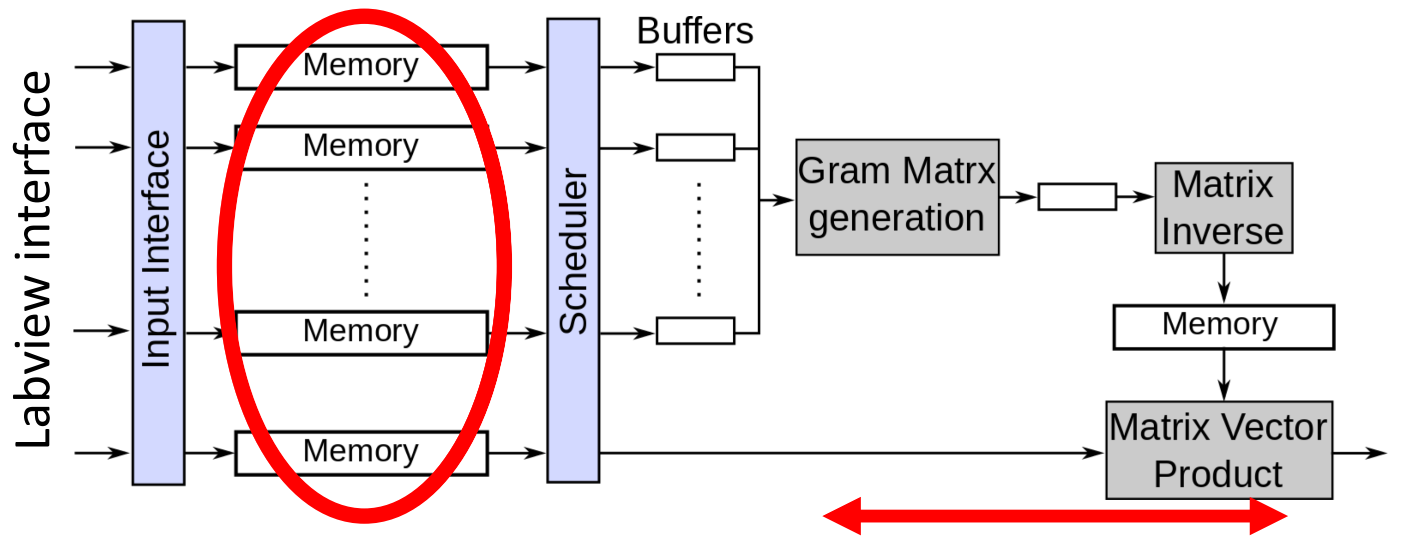
$$H^{\dagger} = H^{H}(HH^{H})^{-1}$$

Inner matrix multiplication – $0.5\, MK^2$
Matrix Inversion – $cK^3$
Matrix-vector multiplication – $N_{int}(K^2 + MK)$

- Needs to be performed frequently (coherence time) and also over the sub-carriers.

- Hence the demand for hardware accelerators

# Linear Detector in LuMAMI




Neumann Series to perform matrix inversion

$$\boldsymbol{Z}^{-1} \approx \sum_{n=0}^{L} \left(\boldsymbol{I}_K - \boldsymbol{X}^{-1}\boldsymbol{Z}\right)^n \boldsymbol{X}^{-1},$$

requires simple matrix multiplications, and has high degree of re-use.

LUND UNIVERSITY

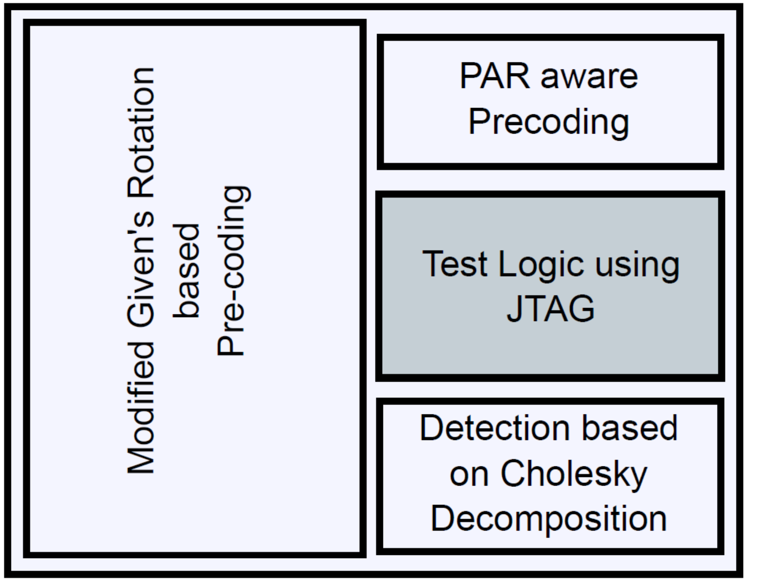
# FPGA Results

- Occupies 270 DSP blocks, 15% of Kintex-7 410T
- Clock frequency of 150 MHz.
- Takes around 0.11 ms to perform detection over 150 sub-carriers.

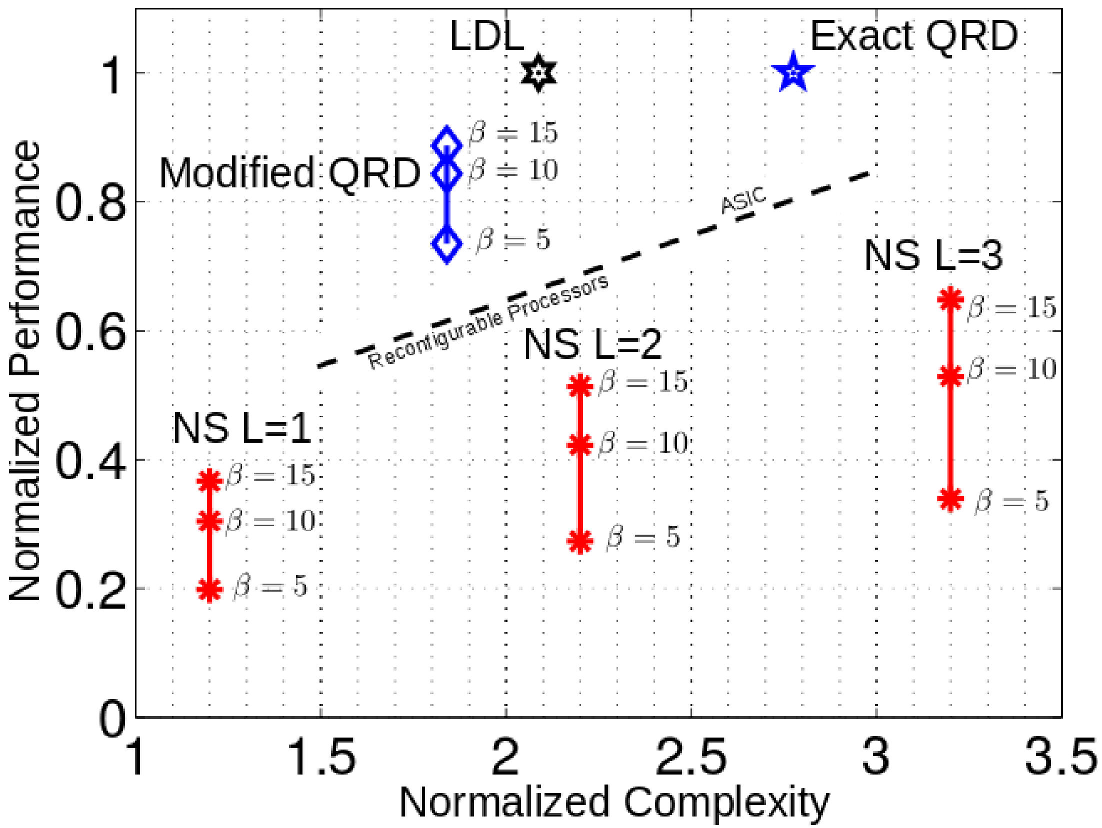
# Pre-coder in ST-28nm

The implementation has 4 modules

- QR decomposition for pre-coding
- Cholesky decomposition for detection
- PAR aware pre-coding
- JTAG based test logic



# Algorithm Evaluation

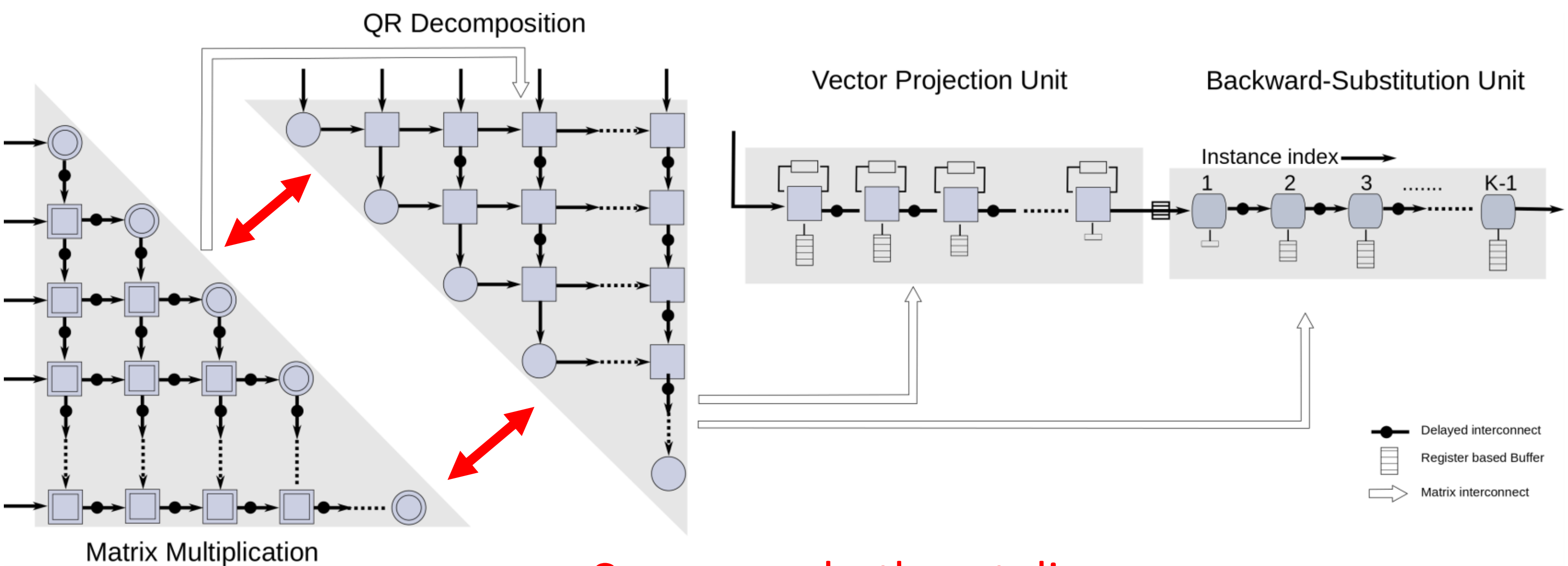



NS – Neumann Series (L – iterations)

LDL – Cholesky Decomposition

Beta is the ratio of number of base station antennas to users.

# Top Level Architecture for pre-coder



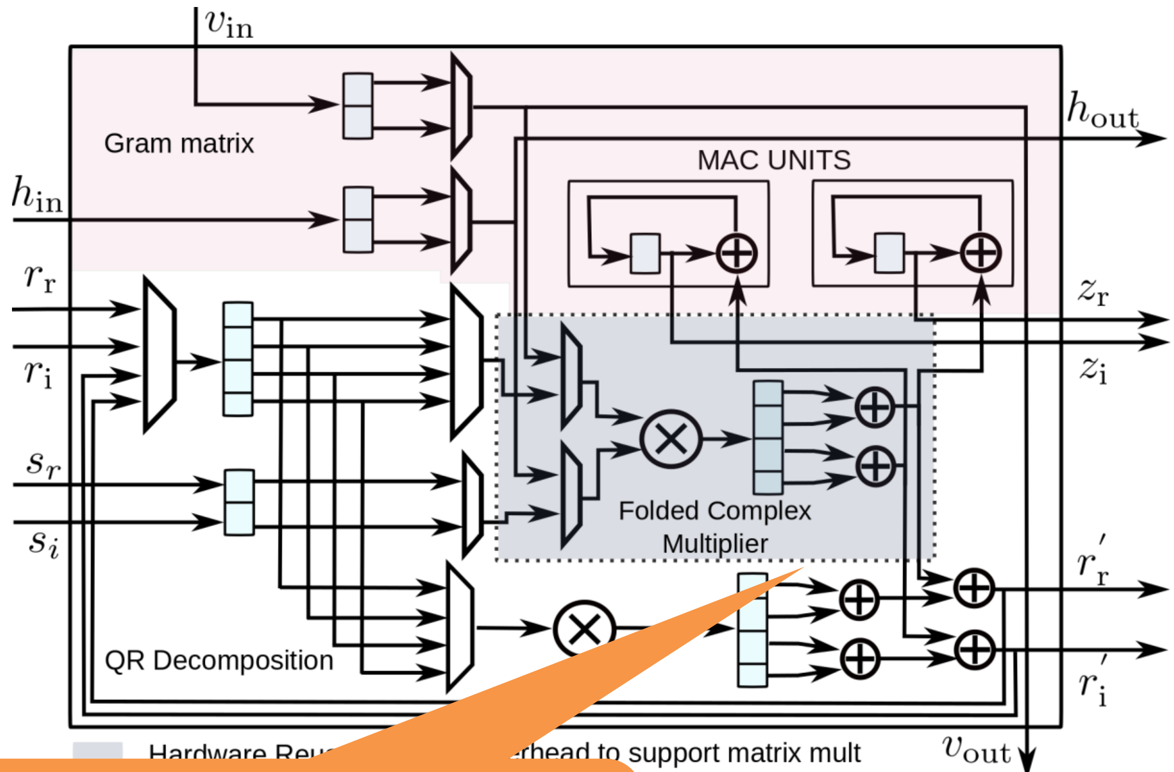

Can merge both systolic arrays

Systolic Arrays --- high throughput, high flexiblity, simple scheduling, and easy design/verification.

Avoid generating Q matrix and inverse of R matrix explicitly.

LUND UNIVERSITY

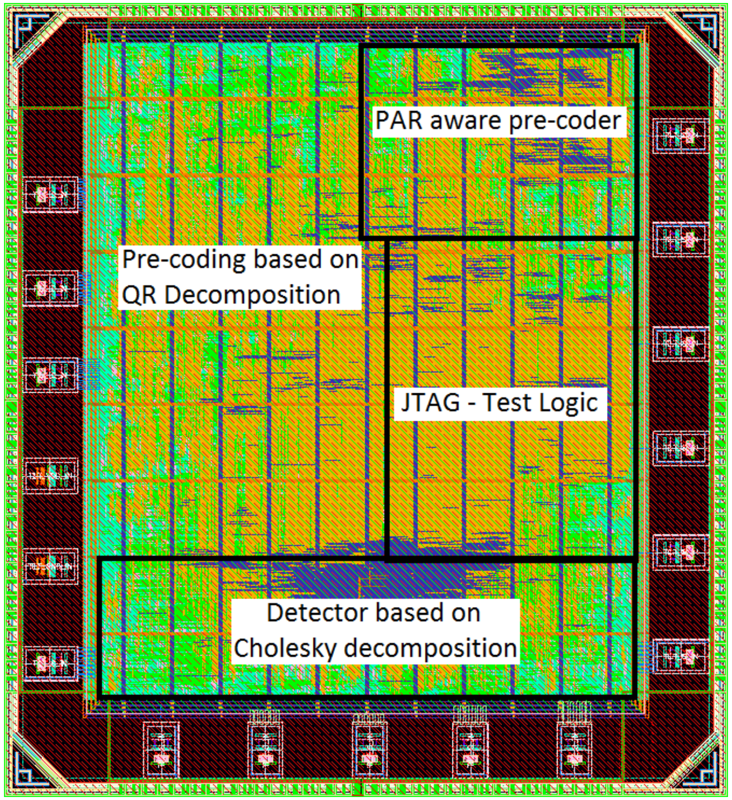
# Unified Processing Element




**Hardware reusability – perform matrix mult and QRD.**
**Highly time-multiplexed**

# ASIC Results

- Supports antenna configurations upto 128x8 (128 base-station antennas and 8 users).
- Total die area of 1mm^2 with max freq of 250MHz at 1 V, and power consumption of 29mW.
- Performance
  - Performs 8x8 QR decomposition in 72 cycles.
  - Performs 8x8 cholesky based data detection in 325 cycles.




ST 28nm FD-SOI implementation

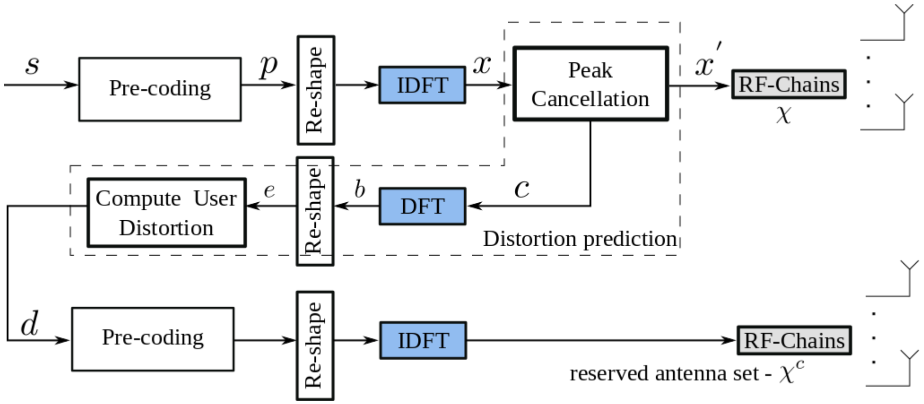
# Pre-coding strategies to tackle PAR

- PAR is a well known problem in OFDM based systems, with techniques like ”tone reservation” to tackle it.
- Massive MIMO inherently has a large degree of freedom (antennas) which can be utilized to reduce PAR.
- One technique we coined as ”Antenna Reservation”
- Constant Envelope pre-coding

LUND UNIVERSITY

# PAR Aware Pre-coding – Antenna Reservation

- Low complexity and existing architecture can be re-used for this technique.
- Lowers back-off by 3 dB with only 15% increase in complexity



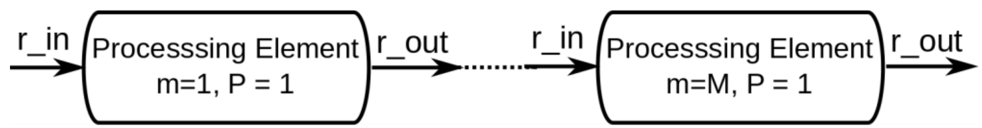
# Constant Envelope pre-coding

Pre-coding also can be seen as

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad ||\alpha_{\text{Tr}}\, \boldsymbol{s} - \boldsymbol{H}\boldsymbol{x}||_2$$

An additional stringent constraint can be added to completly matigate PAR.

This strigent constraint based pre-coding can be solved using cordinate-descent method.




**Each Processing Element**

**Area – 0.03 mm^2**
**Max Freq – 500 MHz**
**Very power hungry – 3.96 mW**

# Conclusion

The matrix properties arising in massive MIMO can be utilized to implement efficient hardware.

Neumann series is very good for implementing fast proto-types on FPGA and for hardware re-use.

The large degree of freedom (antennas)  is exploited to reduce PAR and can be used to tackle other hardware impairments.

LUND UNIVERSITY