Argos: Practical Base Stations for Large-scale Beamforming

Clayton W. Shepard
Collaborators

Hang Yu

Narendra Anand

Erran Li

Thomas Marzetta

Richard Yang

Lin Zhong
Background

• Beamforming
  – Power Gain
  – Adjust phase ("beamweights")
  – Leverages Interference

• Open-loop
  – Pre-compute weights to specify direction

• Closed-loop (adaptive)
  – Use channel state information (CSI) to target receivers
Background

• Single-user beamforming (SUBF)

\[ W_{SUBF} = c \cdot H^* \]

• Multi-user beamforming (MUBF)

\[ W_{MUBF} = c \cdot H^* (H^T H^*)^{-1} \]
Background: Channel Estimation

A pilot is sent from each BS antenna. Channels are not reciprocal.

Due to environment and terminal mobility, estimation has to occur quickly and periodically.

Align the phases at the receiver to ensure constructive interference.

\[ + \]

\[ + \]
MUBF linear pre-coding: downlink
MUBF linear pre-coding: uplink
Our vision
Prior Work

• Large-scale beamforming theory

• Real-world beamforming

• Reciprocal calibration
First large-scale beamforming base station
Overview of contributions

• Scalable architecture

• Internal reciprocity calibration

• Novel fully distributed beamforming method
Can beamforming scale with the number of base station antennas?
Not with current techniques!

• CSI acquisition
  – Typically requires # of base station (BS) antennas (M) + # of terminals (K) pilots

• Weight calculation
  – All existing methods have centralized data dependency
  – Requires M*K channel estimates and produces M*K weight values

• Linear pre-coding
  – Produces M data streams
With careful design and new techniques it can!

- **CSI Acquisition**
  - Leverage TDD reciprocity to limit pilots to K
  - Requires calibration

- **Weight Calculation**
  - Novel decentralized weight calculation

- **Linear Pre-coding**
  - Apply weights at radio
  - For uplink combine streams any time they meet
Scalable linear pre-coding

Common Databus!
MUBF linear pre-coding: uplink
Scalable linear pre-coding

Constant Bandwidth!
Ramifications

• CSI and weights are computed and applied (linear pre-coding) locally at each BS radio
  – No overhead for additional BS radios

• No central data dependency
  – No latency from data transport
  – No stringent latency requirements
  – Constant data rate common bus (no switching!)

• Unlimited scalability!
Design goals

- Scalable
  - Support thousands of BS antennas

- Cost-effective
  - Cost scales linearly with # of antennas

- Reliable
How do we design it?

- **Daisy-chain (series)**
  - Unreliable
  - Large end to end latency

- **Token-ring / Interconnected**
  - Not amenable to linear pre-coding
  - Variable Latency
  - Routing overhead

- **Flat structure**
  - Un-scalable
  - Expensive, with large fixed cost
Solution: Argos

- Modular
  - Daisy-chainable
  - 1 or more radios

- Hierarchal
  - Increases Reliability
  - Constrains Latency
  - Cost-effective
Scalability of Argos

• Scalable in 4 directions:
  – # of Radios per Module
  – # of Modules per Chain
  – # of ports per Hub
  – # of Hubs (and levels)

• Reliable
  – Branches can fail without affecting other branches
  – Central hubs can be easily made redundant

• Accommodates linear pre-coding
  – Add samples together at every junction
Implementation
Overview of contributions

• Scalable architecture

• Internal reciprocity calibration

• Novel fully distributed beamforming method
Channel reciprocity

\[ h_{i \rightarrow j} = tx_i \cdot c \cdot rx_j \]

\[ h_{j \rightarrow i} = tx_j \cdot c \cdot rx_i \]
Calibration coefficients

• Given the complete channel: \( h_{i\rightarrow j} = tx_i \cdot c \cdot rx_j \)

• We define a calibration coefficient as:

\[
A_{i\rightarrow j} = \frac{h_{i\rightarrow j}}{h_{j\rightarrow i}} = \frac{tx_i \cdot c \cdot rx_j}{tx_j \cdot c \cdot rx_i} = \frac{tx_i \cdot rx_j}{tx_j \cdot rx_i} = \frac{1}{A_{j\rightarrow i}}
\]

• Thus:

\[
h_{i\rightarrow j} = A_{i\rightarrow j} h_{j\rightarrow i} \quad \text{and} \quad A_{i\rightarrow j} = \frac{A_{1\rightarrow j}}{A_{1\rightarrow i}}
\]
Applying to large-scale BS

• Find $A$ between each BS antenna and a reference antenna (1) $A_{1\rightarrow m}$

• Every BS radio listens to terminal pilot $h_{t\rightarrow m}$

• Find $A$ between reference and terminal $A_{1\rightarrow t}$

• We can derive $A_{m\rightarrow t} = \frac{A_{1\rightarrow t}}{A_{1\rightarrow m}}$

• Now every $h$ can be found via $h_{m\rightarrow t} = A_{m\rightarrow t}h_{t\rightarrow m}$
Key observation

• But this requires K+1 pilots...
  – Even worse, it requires feedback

• A constant phase shift across the entire array does not alter the beampattern!

\[ h_{m \rightarrow t} = A_{m \rightarrow t} h_{t \rightarrow m} = \frac{A_{1 \rightarrow t}}{A_{1 \rightarrow m}} h_{t \rightarrow m} \Rightarrow \frac{1}{A_{1 \rightarrow m}} h_{t \rightarrow m} \]

• Assuming \( A_{1 \rightarrow t} = 1 \) results in a constant phase offset, and thus does not affect radiation pattern
Internal calibration

• We find all $A_{1\rightarrow m}$ offline
  – They are static, and can be found quickly

• Send K orthogonal pilots to find all $h_{tk\rightarrow m}$
  – Used for uplink beamforming directly

• Use $h_{m\rightarrow t} = \frac{h_{t\rightarrow m}}{A_{1\rightarrow m}}$ for downlink beamforming
Overview of contributions

• Scalable architecture

• Internal reciprocity calibration

• Novel fully distributed beamforming method
Problem with existing methods

• Central data dependency

• Transport latency causes capacity loss

• Can not scale
  – Becomes exorbitantly expensive then infeasible
Conjugate beamforming

• Requires global power scaling by constant:

\[ W_{conj} = c \cdot H^* \]

• Where, e.g.:

\[ c = \left( \sum_{k=1}^{K} \sum_{m=1}^{M} ||h_{m,k}^2|| \right)^{-1} \]

• This creates a central data dependency
Local conjugate beamforming

• Scale power locally:

\[ c_m = \left( \sum_{k=1}^{K} \| h_{m,k}^2 \| \right)^{-1} \quad (m = 1, 2, \ldots, M) \]

• Maximizes utilization of every radio
  – More appropriate for real-world deployments
• Quickly approaches optimal as K increases
  – Channels are independent and uncorrelated
Results

• Huge Capacity Gains

• Performance linear with M and K

• Channel Calibration Stable

• Local conjugate indistinguishable from global
  – Approaches optimality quickly with K
Results: scaling M

Capacity vs. M, with K = 15
Results: scaling K

Capacity vs. K, with M = 64

- Zeroforcing
- Conjugate
- Local Conj
Results: scaling K

Capacity vs. K, with M = 16

- Zeroforcing
- Conjugate
- Local Conj

Total Capacity (bps/Hz) vs. Number of Terminals.
Results: low power

Capacity vs. K, with M = 16

Zero forcing
Conjugate
Local Conj

Number of Terminals
Total Capacity (bps/Hz)
Results: calibration stability
Results: local conjugate
Future directions

• Find optimal tradeoff between zero-forcing and conjugate

• Demonstrate network optimality
  – Lower power reduces other-cell interference
  – Leverage cooperative beamforming

• Investigate promising match with full duplex
  – Leverage huge EIRP gains
Conclusion

• First large-scale beamforming platform
  – Real-world demonstration of manyfold capacity increase

• Devised novel architecture and techniques
  – Unlimited Scalability
Acknowledgements

• Theoretical Discussion and Background
  – Ashutosh Sabharwal

• WARP Support
  – Patrick Murphy, Gaurav Patel, Chris Hunter, Sidharth Gupta

• Platform Construction
  – Nathan Zuege, Chris Harris, Azalia Mirhoseini, Danny Eaton, Paul Williams