PREDICTING COMMUNICATION PERFORMANCE

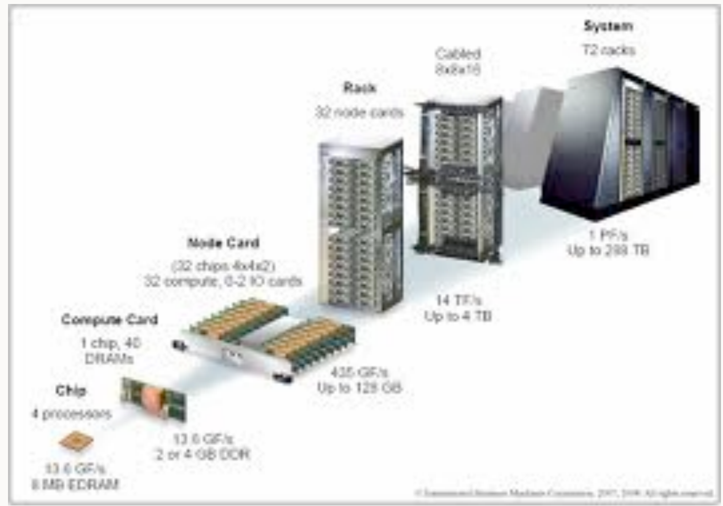
Nikhil Jain
CASC Seminar, LLNL

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.
This work was funded by the Laboratory Directed Research and Development Program at LLNL under project tracking code 13-ERD-055

SUPERCOMPUTERS

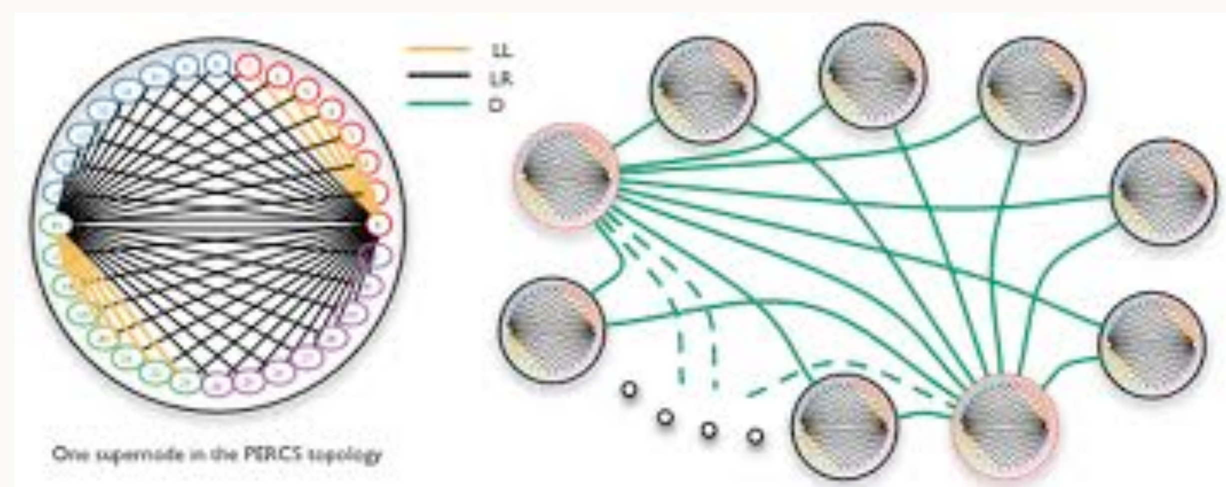
SUPERCOMPUTERS

48 GB/s, 1-2 microsec



40 GB/s, 1-3 microsec

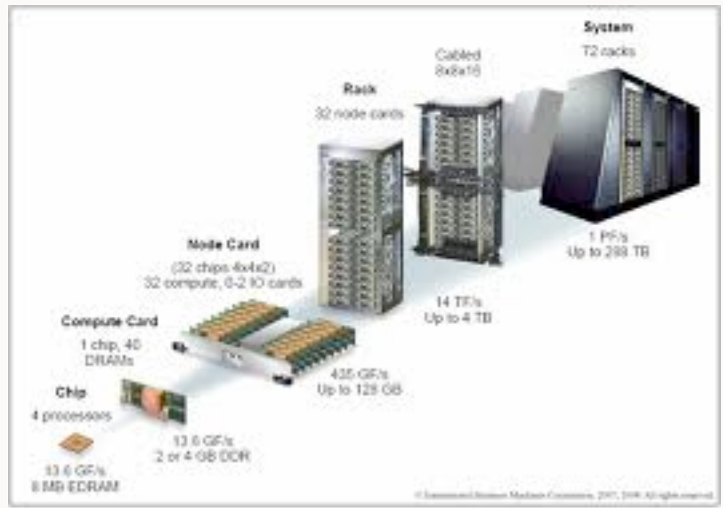
150 GB/s, 0.8 microsec



420 GB/s, 1-2 microsec

SUPERCOMPUTERS

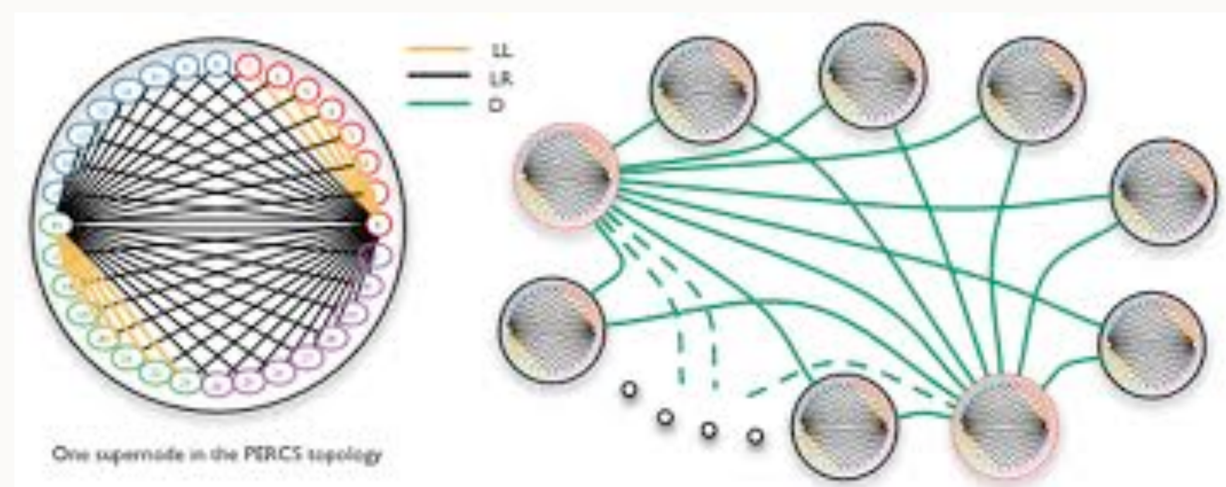
48 GB/s, 1-2 microsec



40 GB/s, 1-3 microsec

150 GB/s, 0.8 microsec

Larger Bandwidth
Lower Latency
Fewer hops



420 GB/s, 1-2 microsec

WHY STUDY NETWORK PERFORMANCE?

WHY STUDY NETWORK PERFORMANCE?

- Peak bandwidth and latency are never realized in presence of congestion

WHY STUDY NETWORK PERFORMANCE?

- Peak bandwidth and latency are never realized in presence of congestion
- High raw bandwidth does not guarantee proportionate observed performance
  - Blue Gene vs Cray's Gemini

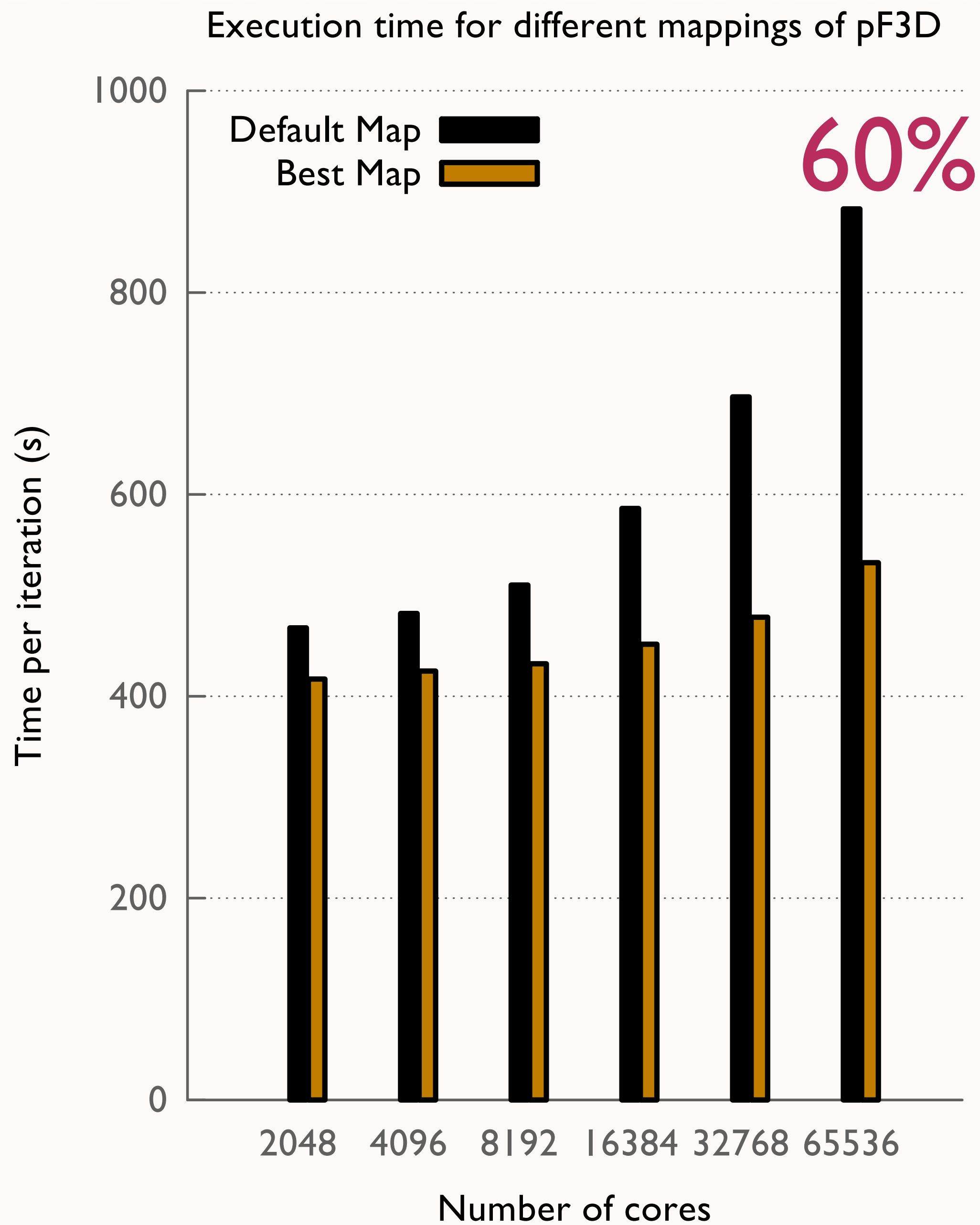
WHY STUDY NETWORK PERFORMANCE?

- Peak bandwidth and latency are never realized in presence of congestion
- High raw bandwidth does not guarantee proportionate observed performance
  - Blue Gene vs Cray's Gemini
- Savings are proportionate to core-count

WHY STUDY NETWORK PERFORMANCE?

- Peak bandwidth and latency are never realized in presence of congestion
- High raw bandwidth does not guarantee proportionate observed performance
  - Blue Gene vs Cray's Gemini
- Savings are proportionate to core-count
- Most importantly, as a graduate student, I do what I am asked to do!

QUANTIFYING IMPACT



- Mapping via logical operations in Rubik
- What about others mappings?
- How far are we from the best?

A. Bhatele, et al Mapping applications with collectives over sub-communicators on torus networks. In Proceedings of the ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis, SC '12. IEEE Computer Society, Nov. 2012 (to appear). LLNL-CONF-556491.

ALTERNATIVES

*Abhinav Bhatele, Nikhil Jain, William D. Gropp, and Laxmikant V. Kale. 2011b. Avoiding hot-spots on two- level direct networks. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC '11)*. ACM, New York, NY, USA, 76:1–76:11.

ALTERNATIVES

- Theoretically: NP hard
- Simulations: too slow
  - 15 days to simulate one use case*

*Abhinav Bhatele, Nikhil Jain, William D. Gropp, and Laxmikant V. Kale. 2011b. Avoiding hot-spots on two- level direct networks. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC '11)*. ACM, New York, NY, USA, 76:1–76:11.

ALTERNATIVES

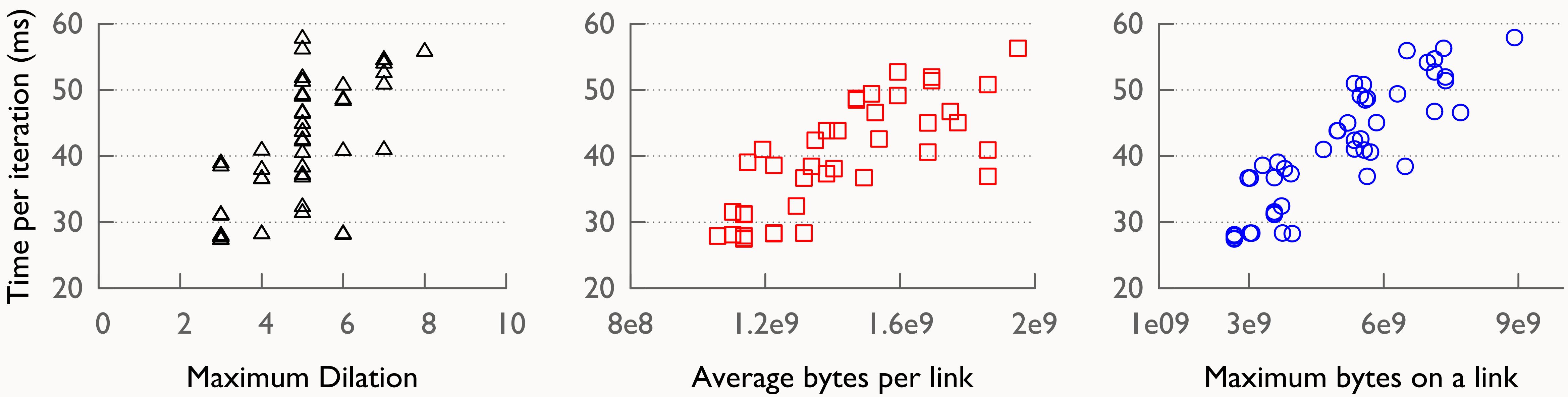
- Theoretically: NP hard
- Simulations: too slow
  - 15 days to simulate one use case*
- Real runs: very expensive
  - Application/allocation specific information

	2012	2013
Intrepid	4.16M	0.73M
Mira	0.17M	7.67M
Total	4.33M	8.40M

13 million core hours!

*Abhinav Bhatele, Nikhil Jain, William D. Gropp, and Laxmikant V. Kale. 2011b. Avoiding hot-spots on two- level direct networks. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC '11)*. ACM, New York, NY, USA, 76:1–76:11.

HEURISTICS - KNOWN METRICS



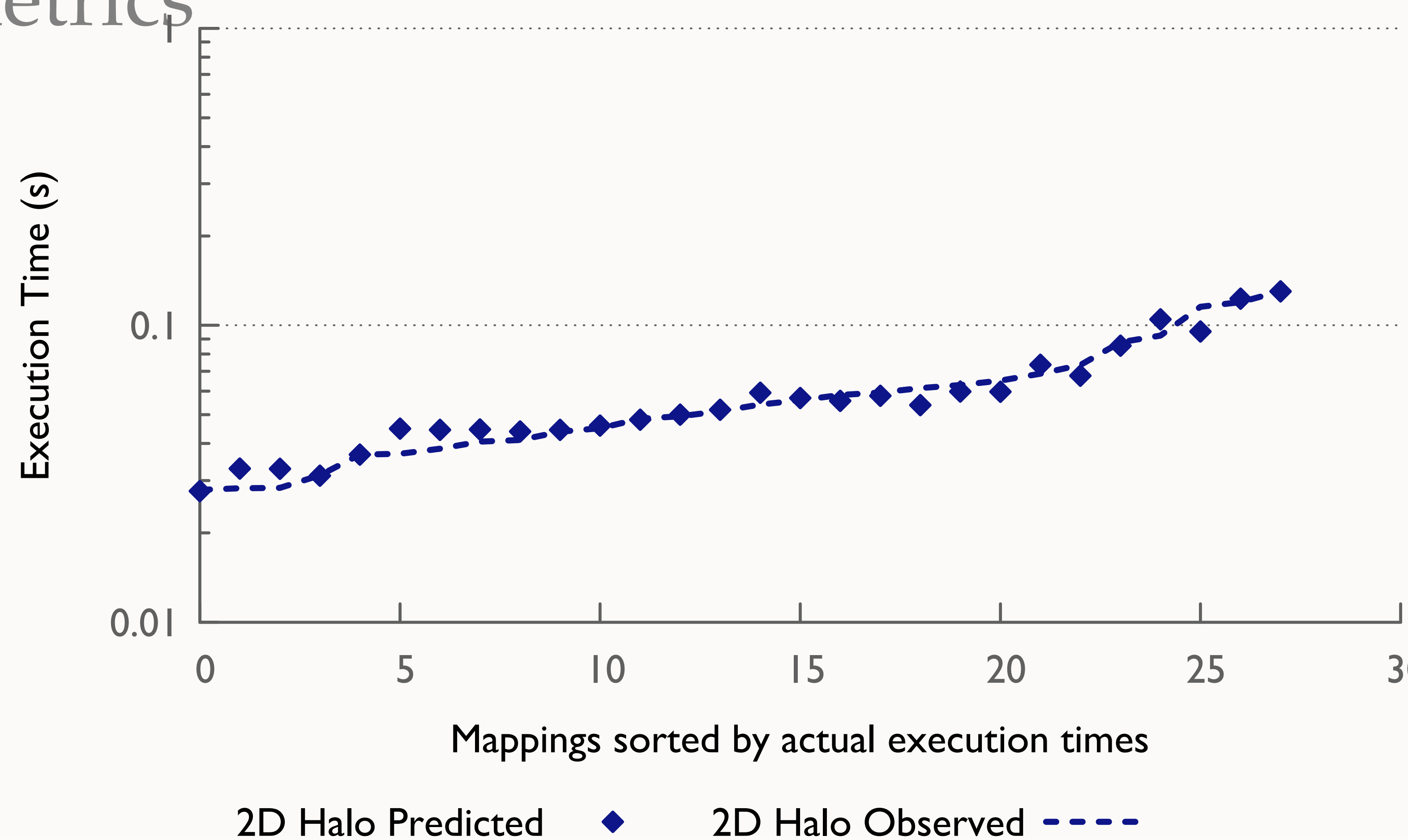
2D-Halo: predicting performance using a linear regression model for known metrics

SUPERVISED LEARNING

- Collect/generate data and summarize
- Build models: train performance prediction based on independent metrics
- Predict

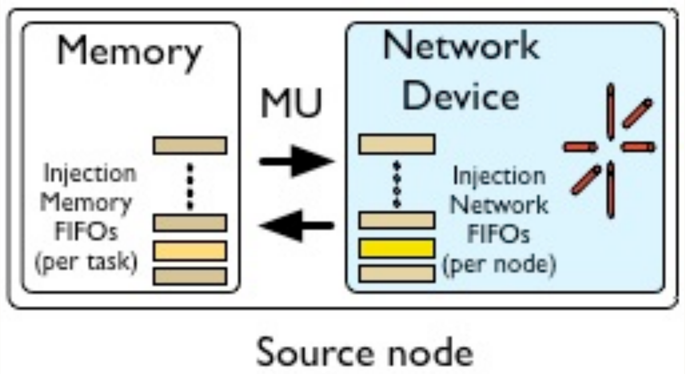
SUPERVISED LEARNING

- Collect/generate data and summarize
- Build models: train performance prediction based on independent metrics
- Predict

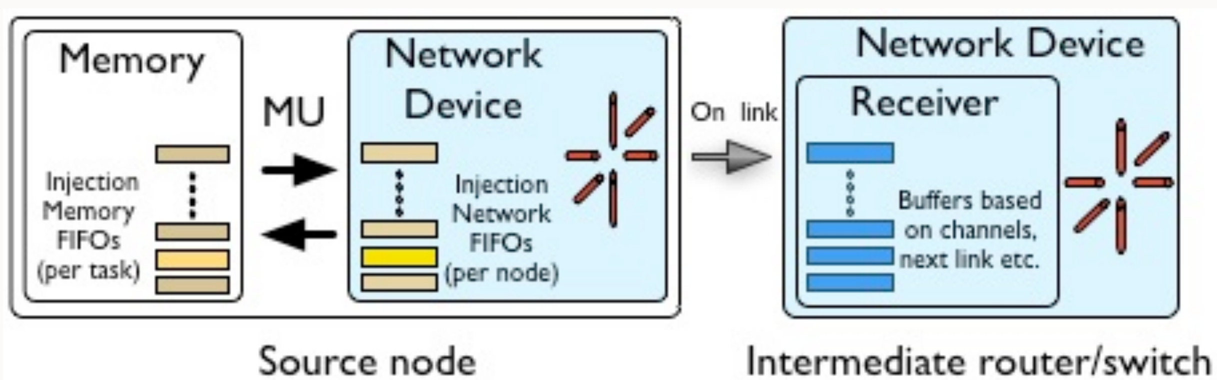


COMMUNICATION

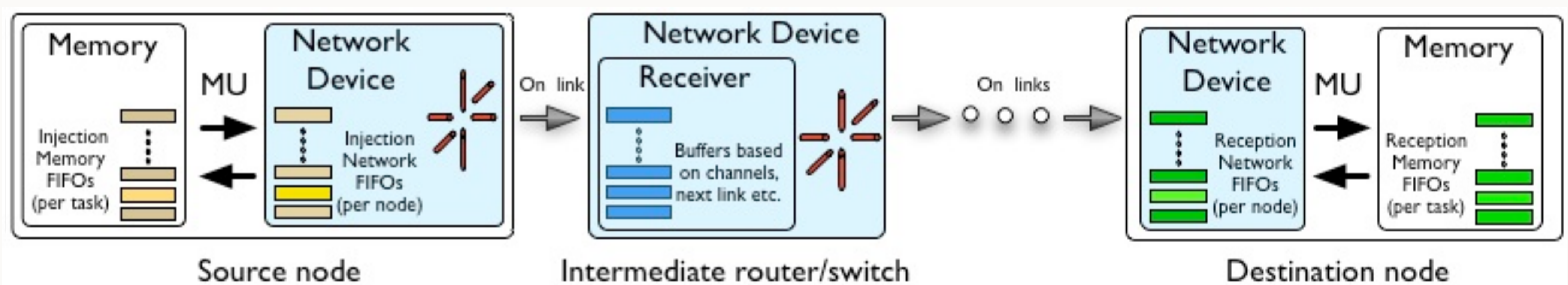
COMMUNICATION



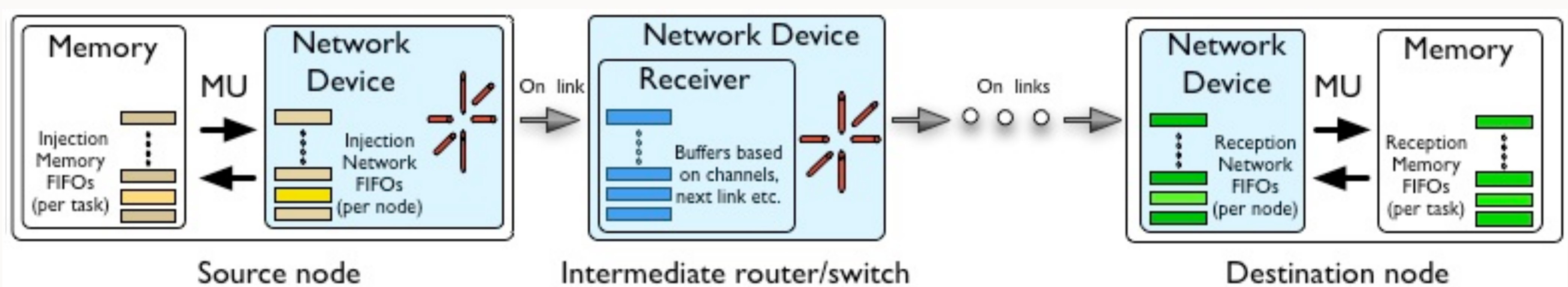
COMMUNICATION



COMMUNICATION

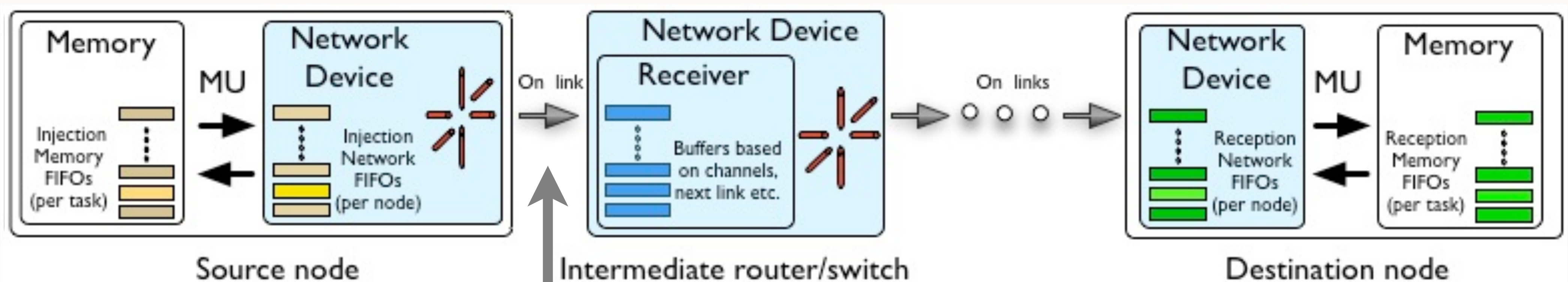


COMMUNICATION



Injection FIFO
Contention

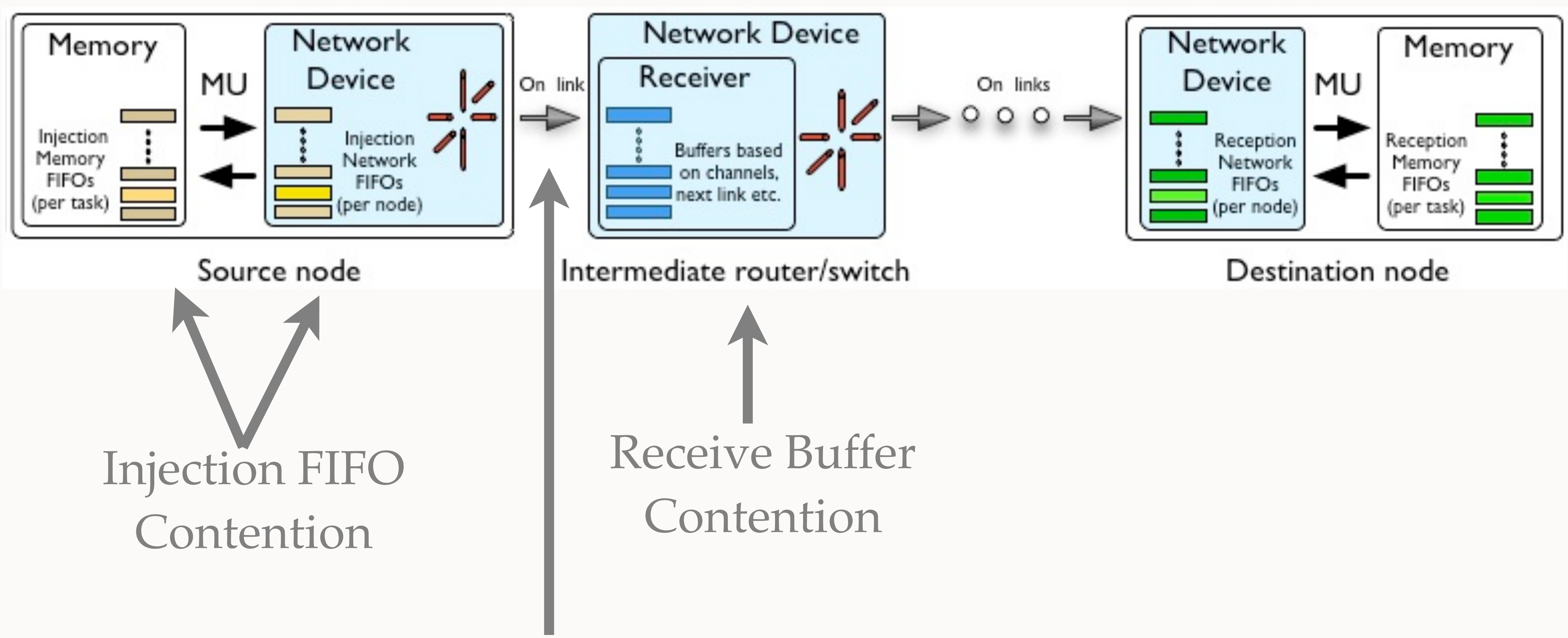
COMMUNICATION



Injection FIFO Contention

Link Contention

COMMUNICATION

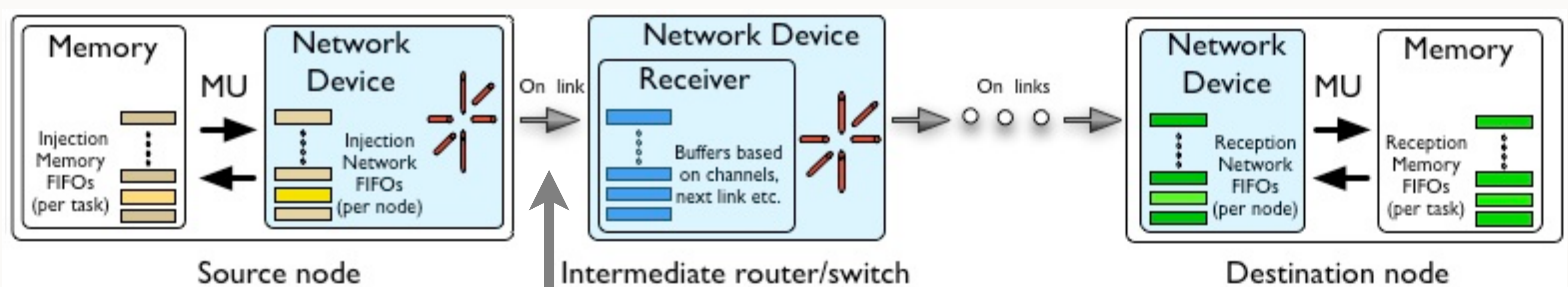


Injection FIFO Contention

Receive Buffer Contention

Link Contention

COMMUNICATION



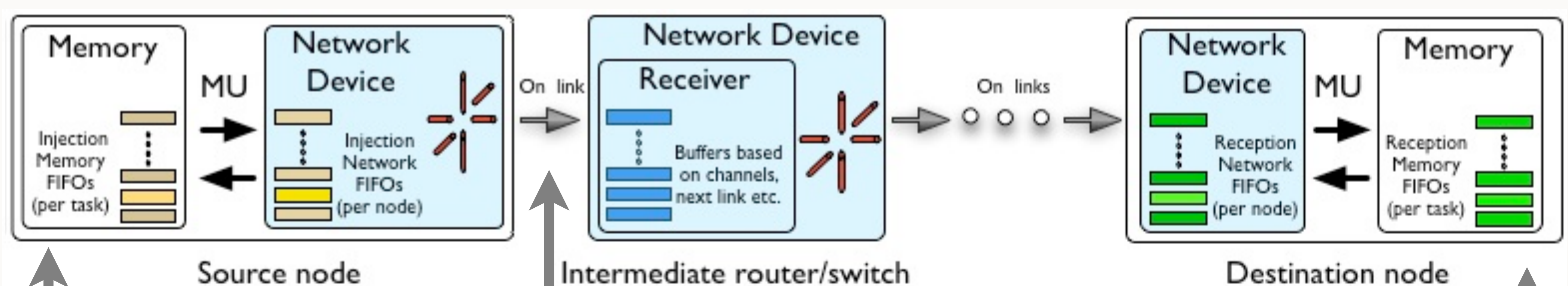
Injection FIFO Contention

Receive Buffer Contention

Reception FIFO Contention

Link Contention

COMMUNICATION



Injection FIFO Contention

Receive Buffer Contention

Reception FIFO Contention

Memory Contention

Link Contention

Memory Contention

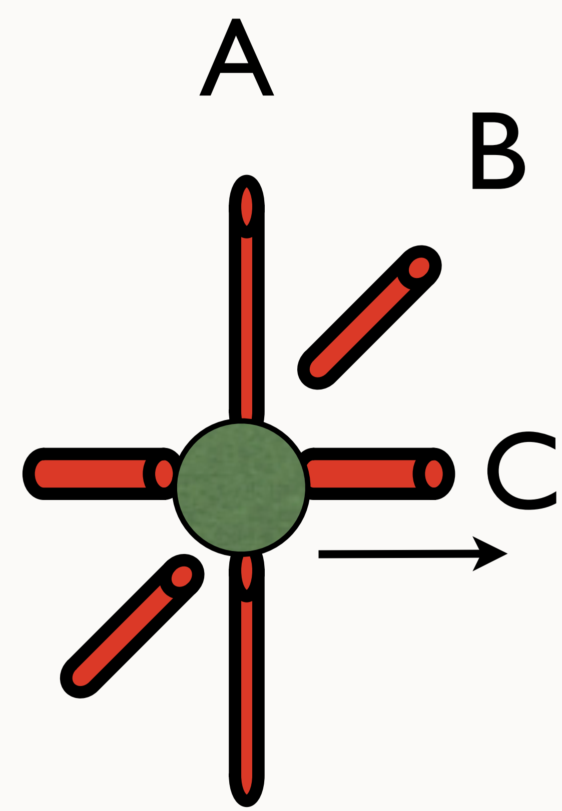
NETWORK COUNTERS OF BLUEGENE/Q

NETWORK COUNTERS OF BLUEGENE/Q

- A PMPI based BGQ-Counter collection module

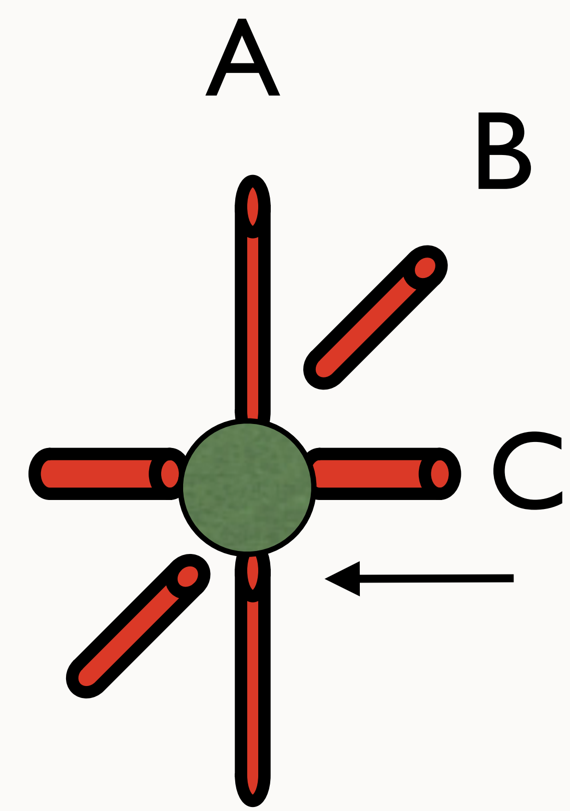
NETWORK COUNTERS OF BLUEGENE/Q

- A PMPI based BGQ-Counter collection module
- Packets sent on links in specific directions: A, B, C, D, E
  - deterministic, dynamic



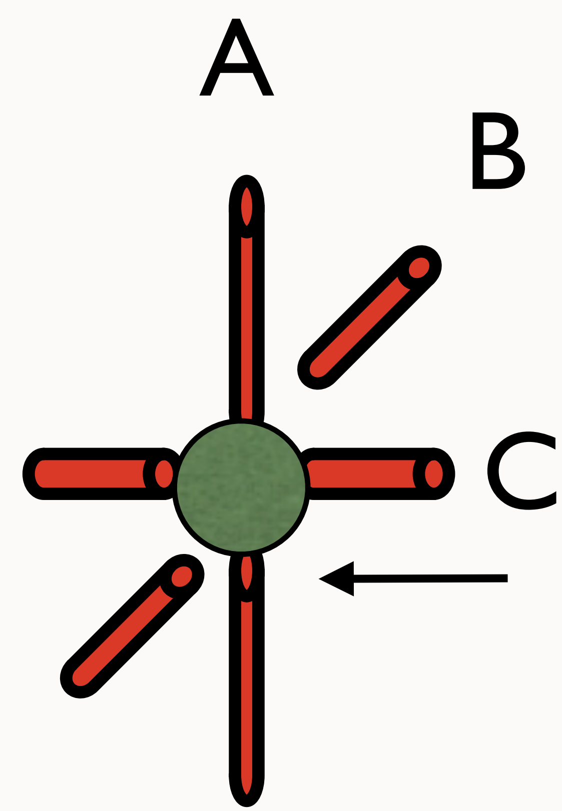
NETWORK COUNTERS OF BLUEGENE/Q

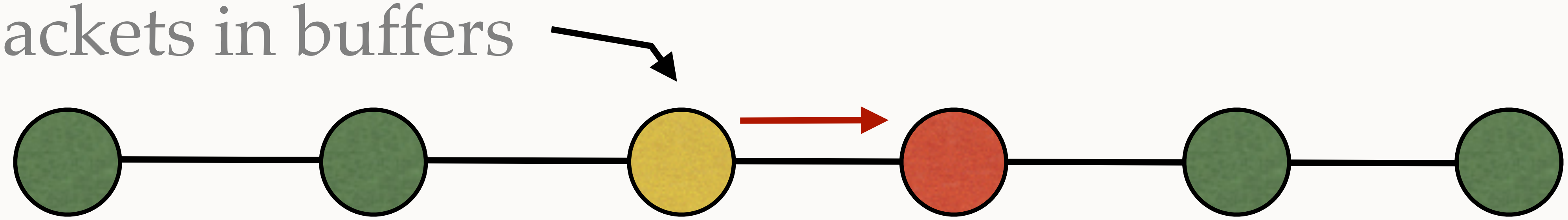
- A PMPI based BGQ-Counter collection module
- Packets sent on links in specific directions: A, B, C, D, E
  - deterministic, dynamic
- Packets received on a link



NETWORK COUNTERS OF BLUEGENE/Q

- A PMPI based BGQ-Counter collection module
- Packets sent on links in specific directions: A, B, C, D, E
  - deterministic, dynamic
- Packets received on a link
- Packets in buffers





ANALYTICAL TOOL

ANALYTICAL TOOL

- Simulate the injection mechanism
    - Selection of memory injection FIFO
    - Mapping of memory FIFO to network injection FIFO

ANALYTICAL TOOL

- Simulate the injection mechanism
  - Selection of memory injection FIFO
  - Mapping of memory FIFO to network injection FIFO
- Simulate routing to obtain hops/dilation

SUPERVISED LEARNING

SUPERVISED LEARNING

- Collect raw data - various entities, e.g. bytes on a link, and the observed performance.

SUPERVISED LEARNING

- Collect raw data - various entities, e.g. bytes on a link, and the observed performance.
- Derive metrics from the raw data on entities, e.g. average of bytes on links.

SUPERVISED LEARNING

- Collect raw data - various entities, e.g. bytes on a link, and the observed performance.
- Derive metrics from the raw data on entities, e.g. average of bytes on links.
- Create a database of derived metrics and performance; we have used 100 mappings.

SUPERVISED LEARNING

- Collect raw data - various entities, e.g. bytes on a link, and the observed performance.
- Derive metrics from the raw data on entities, e.g. average of bytes on links.
- Create a database of derived metrics and performance; we have used 100 mappings.
- Select two-third entries as training set; includes derived metrics and performance.

SUPERVISED LEARNING

SUPERVISED LEARNING

- The training set is used to create a model for prediction

12

SUPERVISED LEARNING

- The training set is used to create a model for prediction
- Remaining entries from the database are used as the test set - only derived metrics.

SUPERVISED LEARNING

- The training set is used to create a model for prediction
- Remaining entries from the database are used as the test set - only derived metrics.
- Prediction is compared with observed values.

SUPERVISED LEARNING

- The training set is used to create a model for prediction
- Remaining entries from the database are used as the test set - only derived metrics.
- Prediction is compared with observed values.
- Experimented with a large number of algorithms - linear, bayesian, SVM, near-neighbors etc.

SUPERVISED LEARNING

- The training set is used to create a model for prediction
- Remaining entries from the database are used as the test set - only derived metrics.
- Prediction is compared with observed values.
- Experimented with a large number of algorithms - linear, bayesian, SVM, near-neighbors etc.

SUPERVISED LEARNING

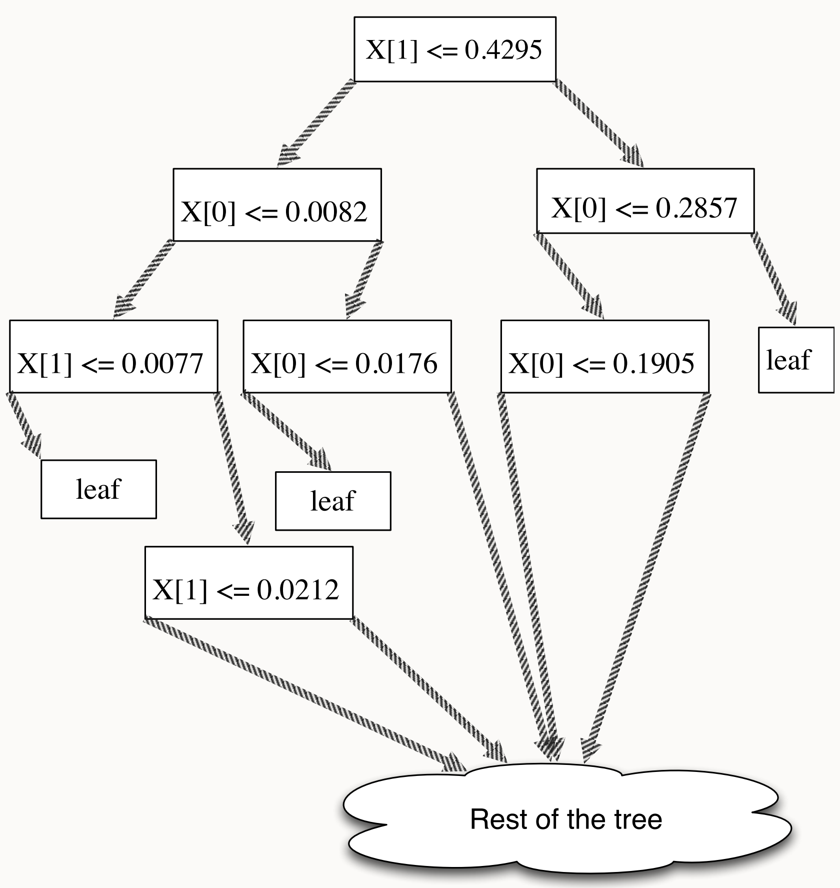
- The training set is used to create a model for prediction
- Remaining entries from the database are used as the test set - only derived metrics.
- Prediction is compared with observed values.
- Experimented with a large number of algorithms - linear, bayesian, SVM, near-neighbors etc.

http://scikit-learn.org

SUPERVISED LEARNING

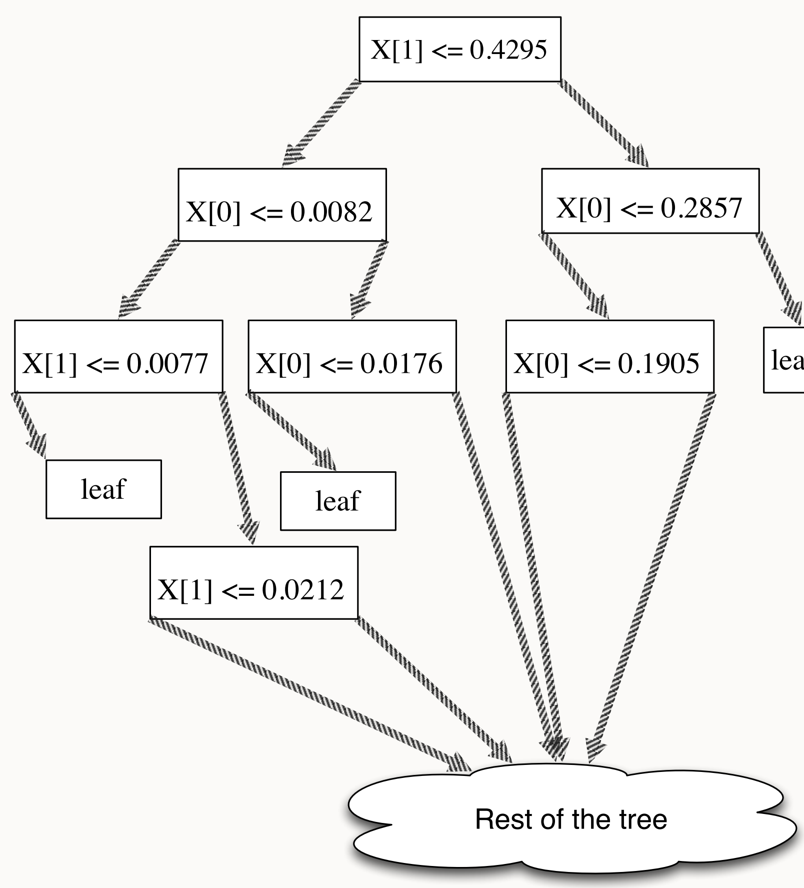
SUPERVISED LEARNING

Decision trees

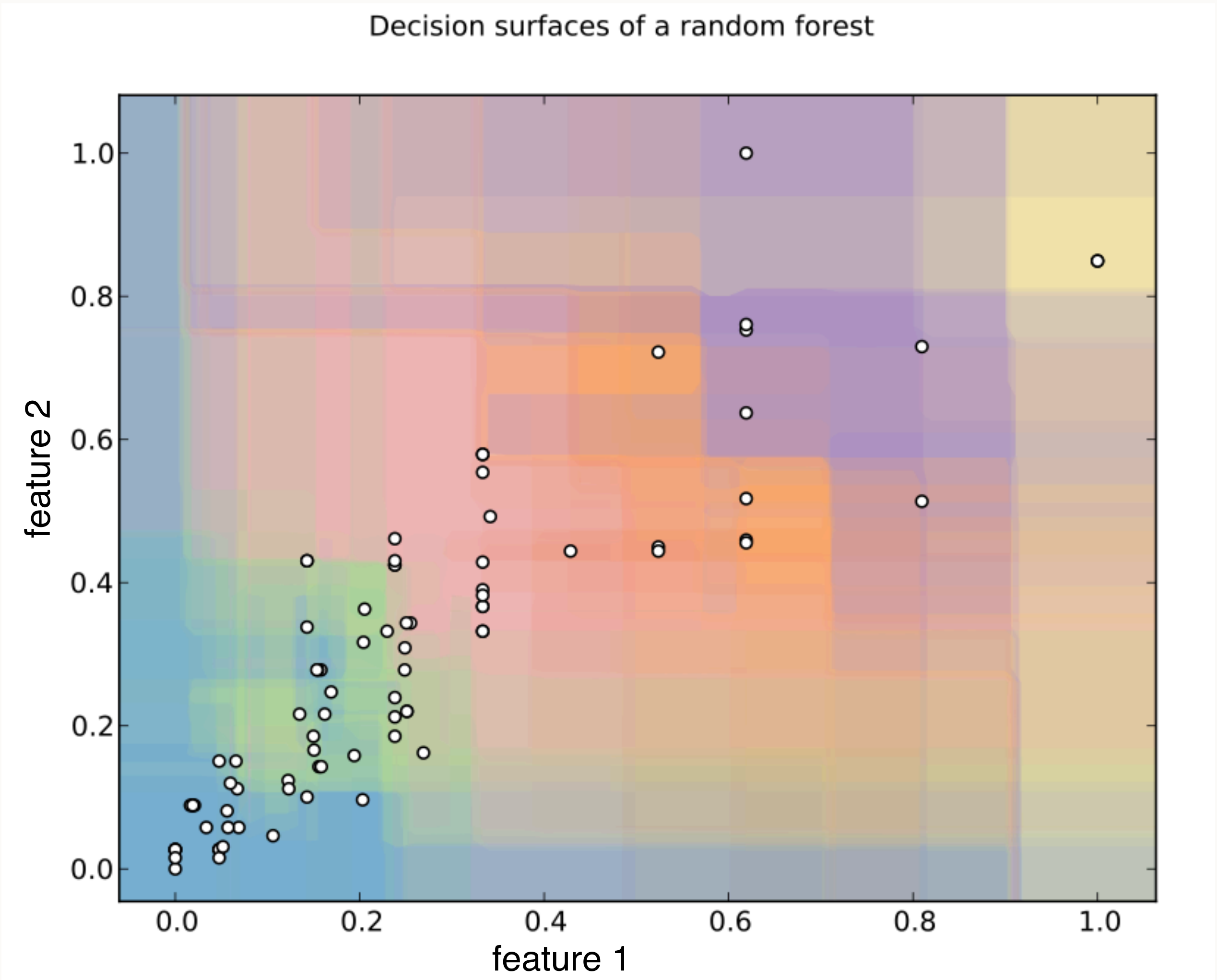


SUPERVISED LEARNING

Decision trees



Randomized forest of trees



HOW TO JUDGE A PREDICTION

- Rank Correlation Coefficient (RCC): fraction of the number of pairs of task mappings whose ranks are in the same partial order in predicted and observed performance list

$$concord_{ij} = \begin{cases} 1, & \text{if } x_i >= x_j \ \& \ y_i >= y_j \\ 1, & \text{if } x_i < x_j \ \& \ y_i < y_j \\ 0, & \text{otherwise} \end{cases}$$

$$RCC = \Big(\sum_{0<=i<n} \sum_{0<=j<i} concord_{ij} \Big) \ /(\frac{n(n-1)}{2})$$

- Absolute Correlation

$$R^2(y, \hat{y}) = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

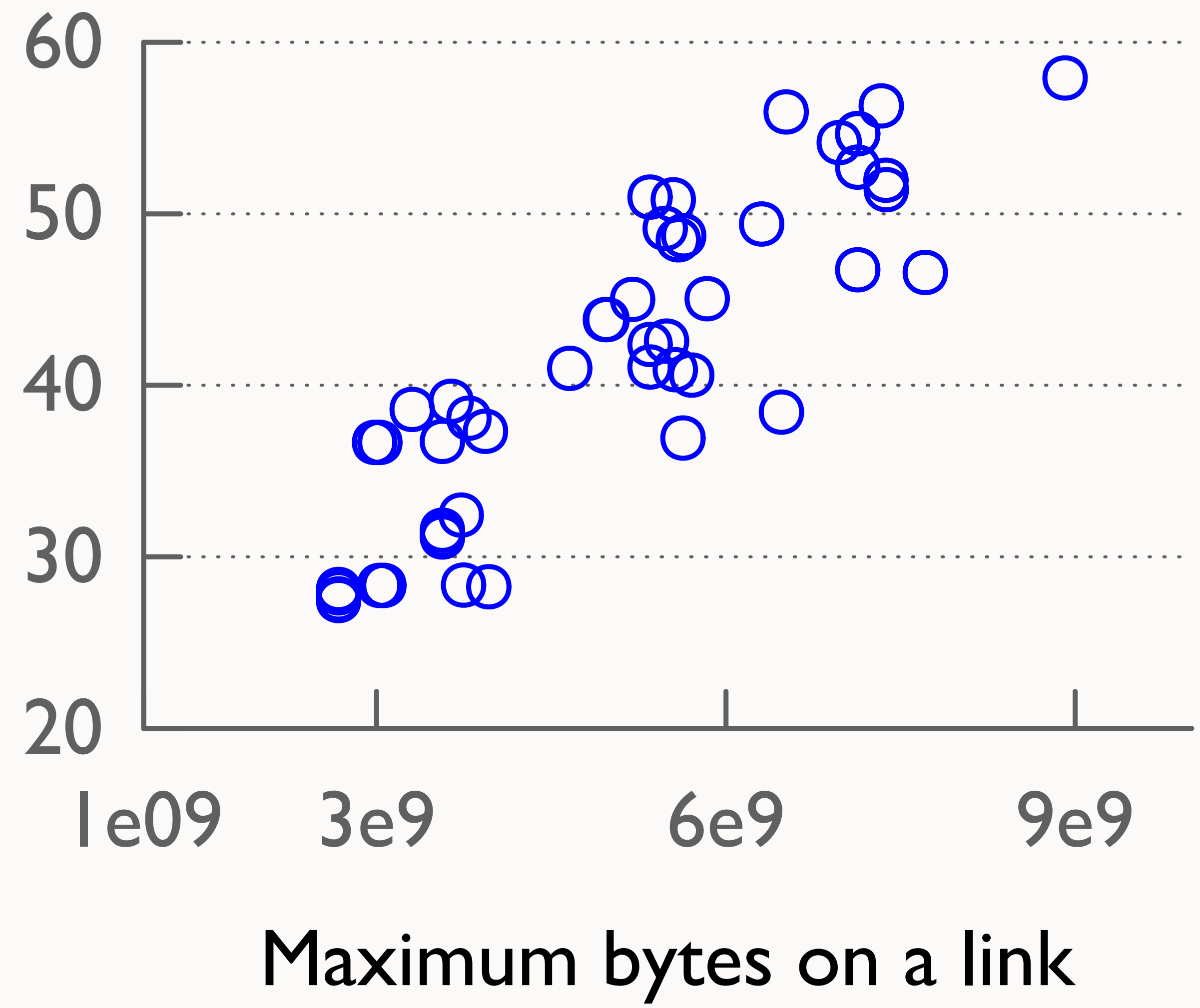
- Higher the better!

RESULTS

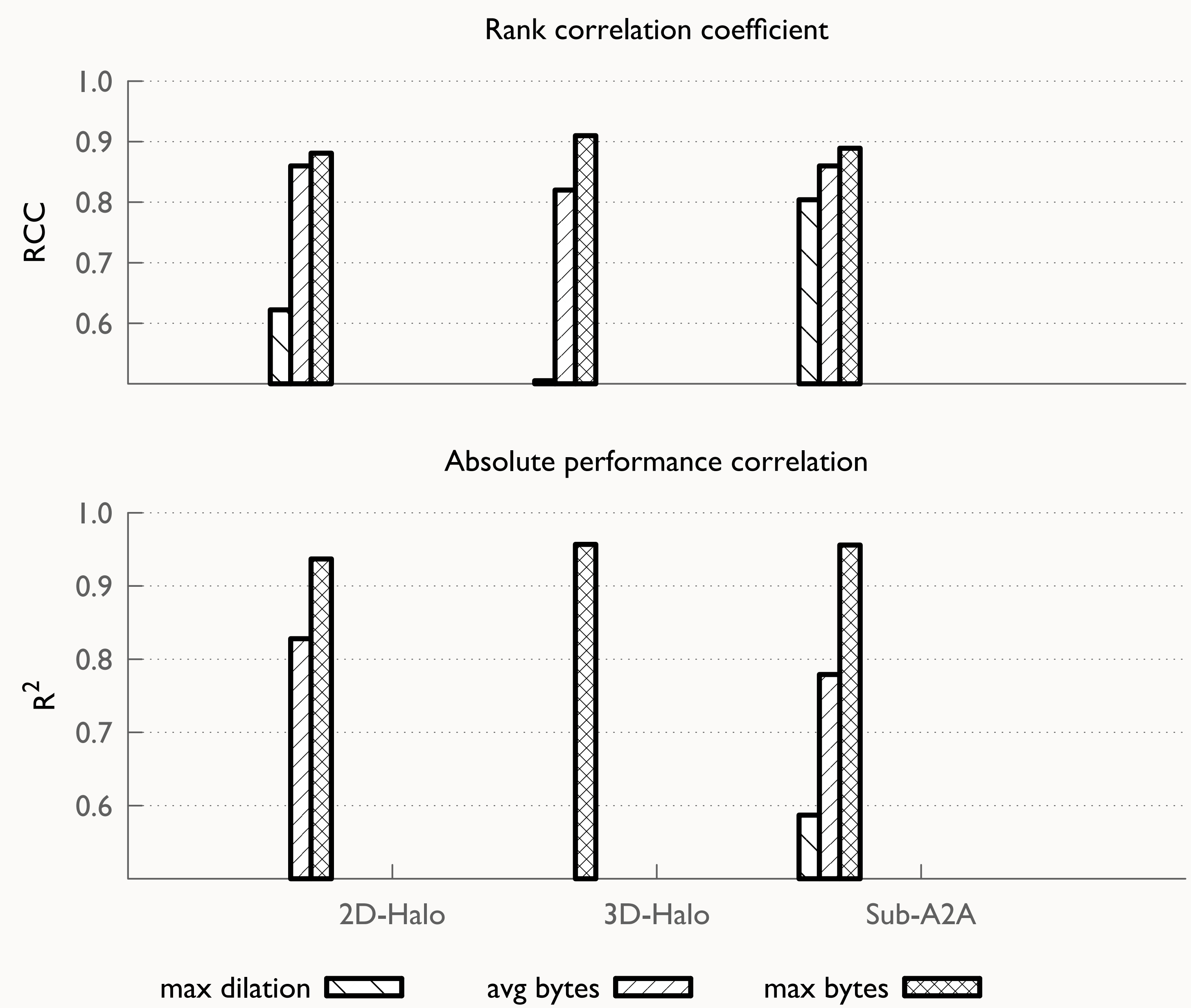
- Three communication kernel
  - Five-point 2D stencil
  - 14-point 3D stencil
  - Sub-communicator all-to-all
- Four message sizes to span MPI and routing protocols

KNOWN METRICS

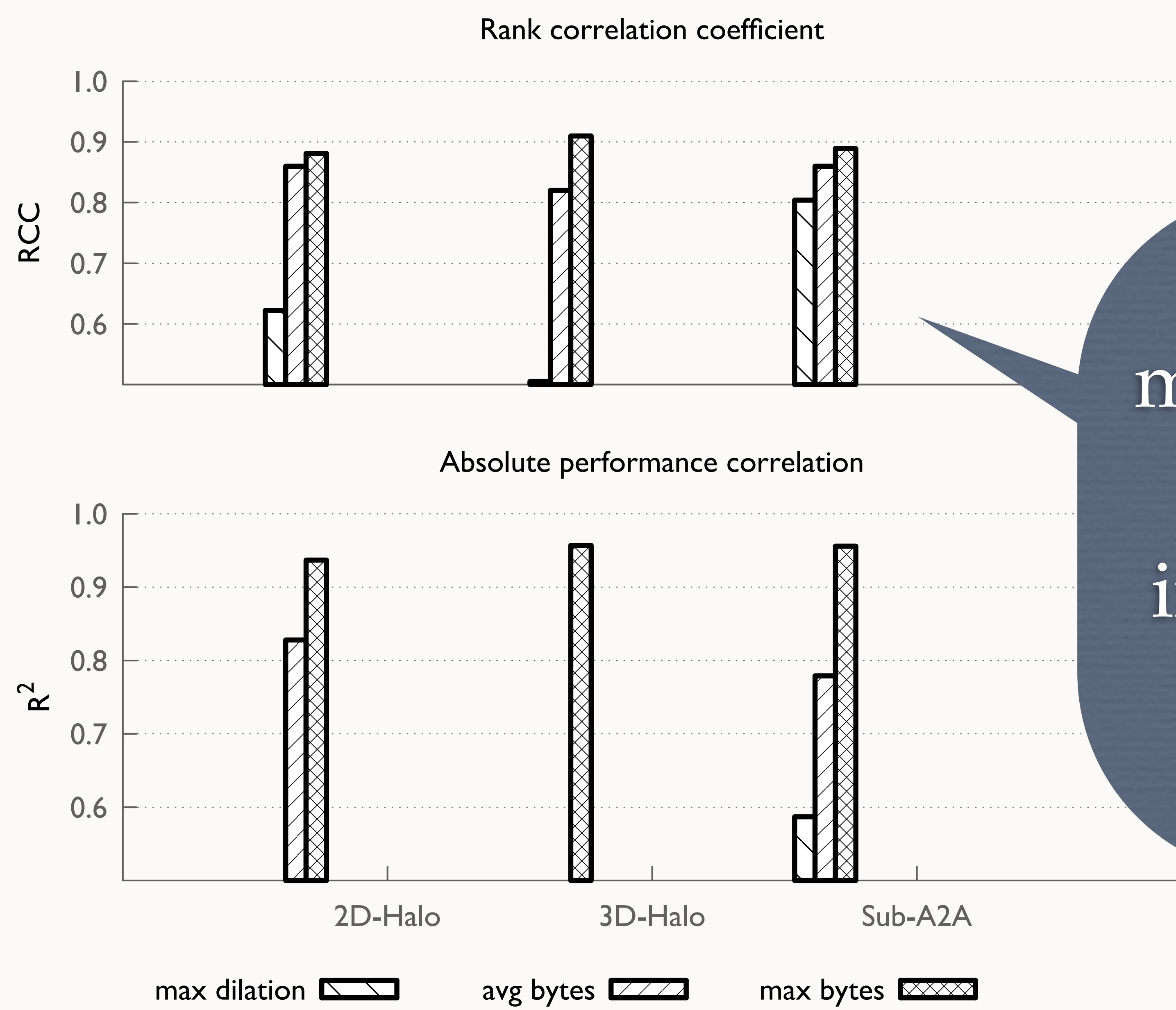
- Entities
  - Bytes on a link
  - Dilation
- Derivation Methods
  - Maximum
  - Average
  - Sum



RESULTS
KNOWN METRICS



RESULTS
KNOWN METRICS



max bytes is good, but incorrect in 10% cases

NEW METRICS

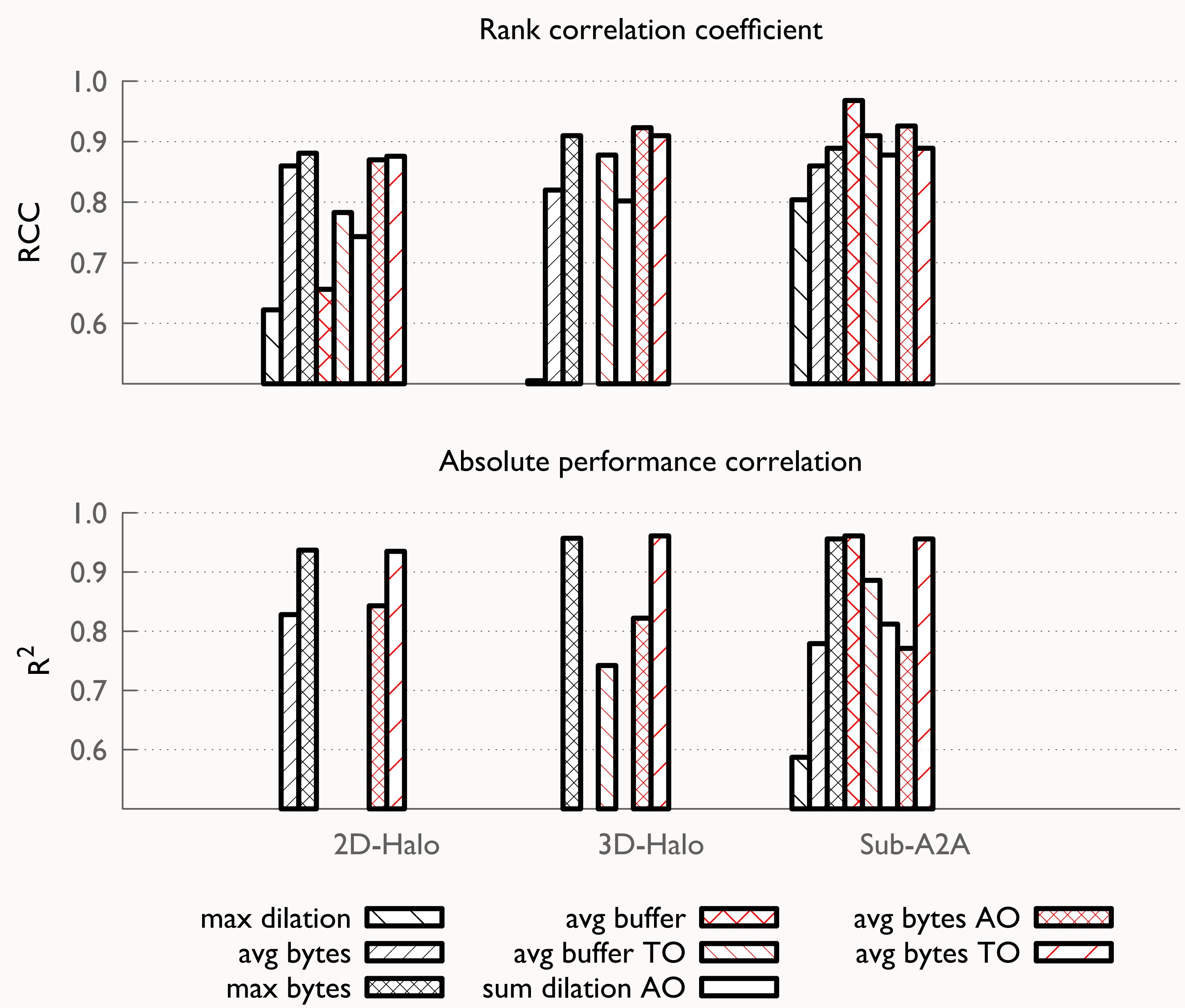
NEW METRICS

- Entities
  - Buffer length (on intermediate nodes)
  - FIFO length (packets in injection FIFO)
  - Delay per link (packets in buffer/packets received)

NEW METRICS

- Entities
  - Buffer length (on intermediate nodes)
  - FIFO length (packets in injection FIFO)
  - Delay per link (packets in buffer/packets received)
- Derivation methods
  - Average Outliers (AO)
  - Top Outliers (TO)

RESULTS
NEW METRICS



HYBRID METRICS

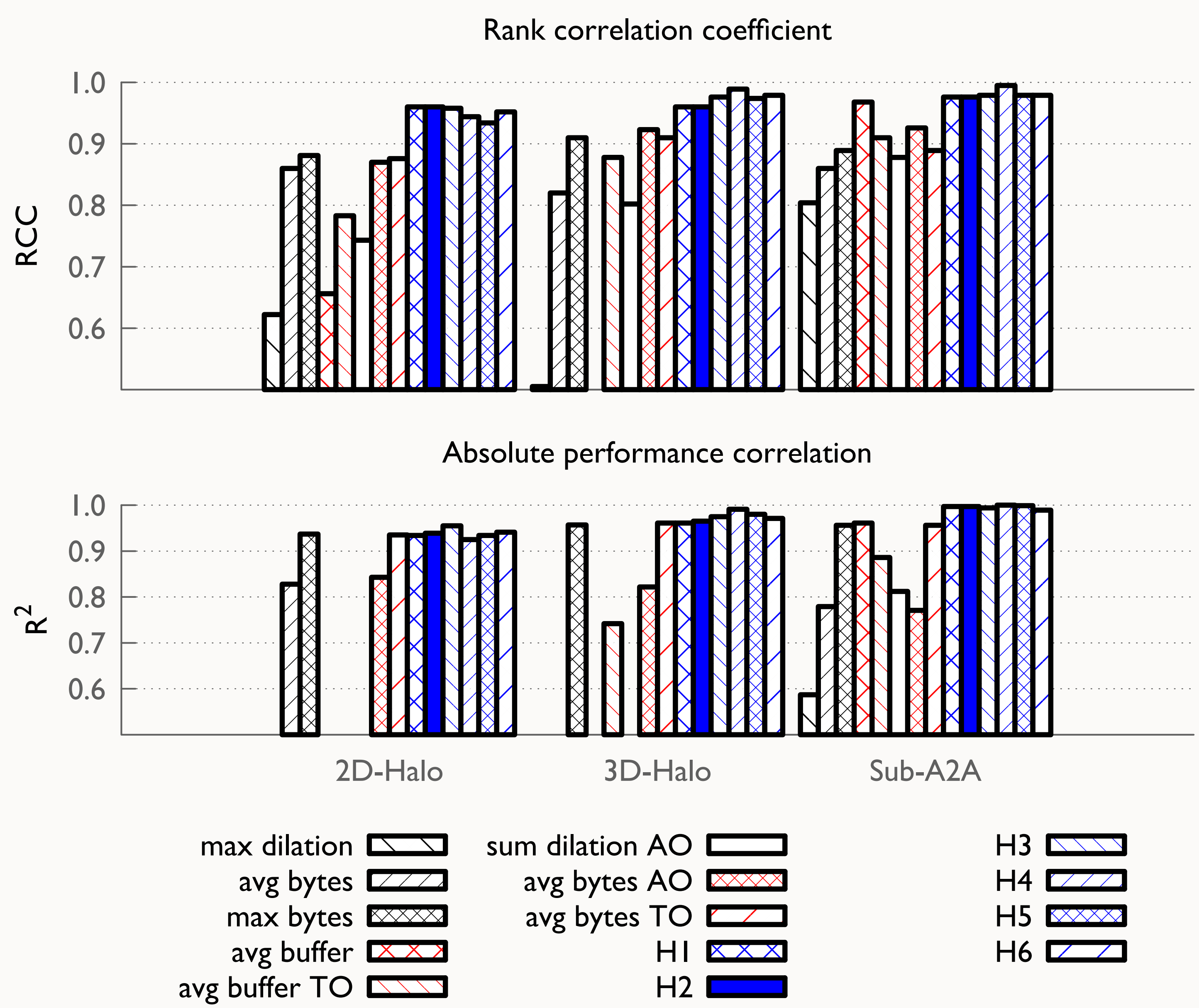
HYBRID METRICS

- Combine multiple metrics to complement each other

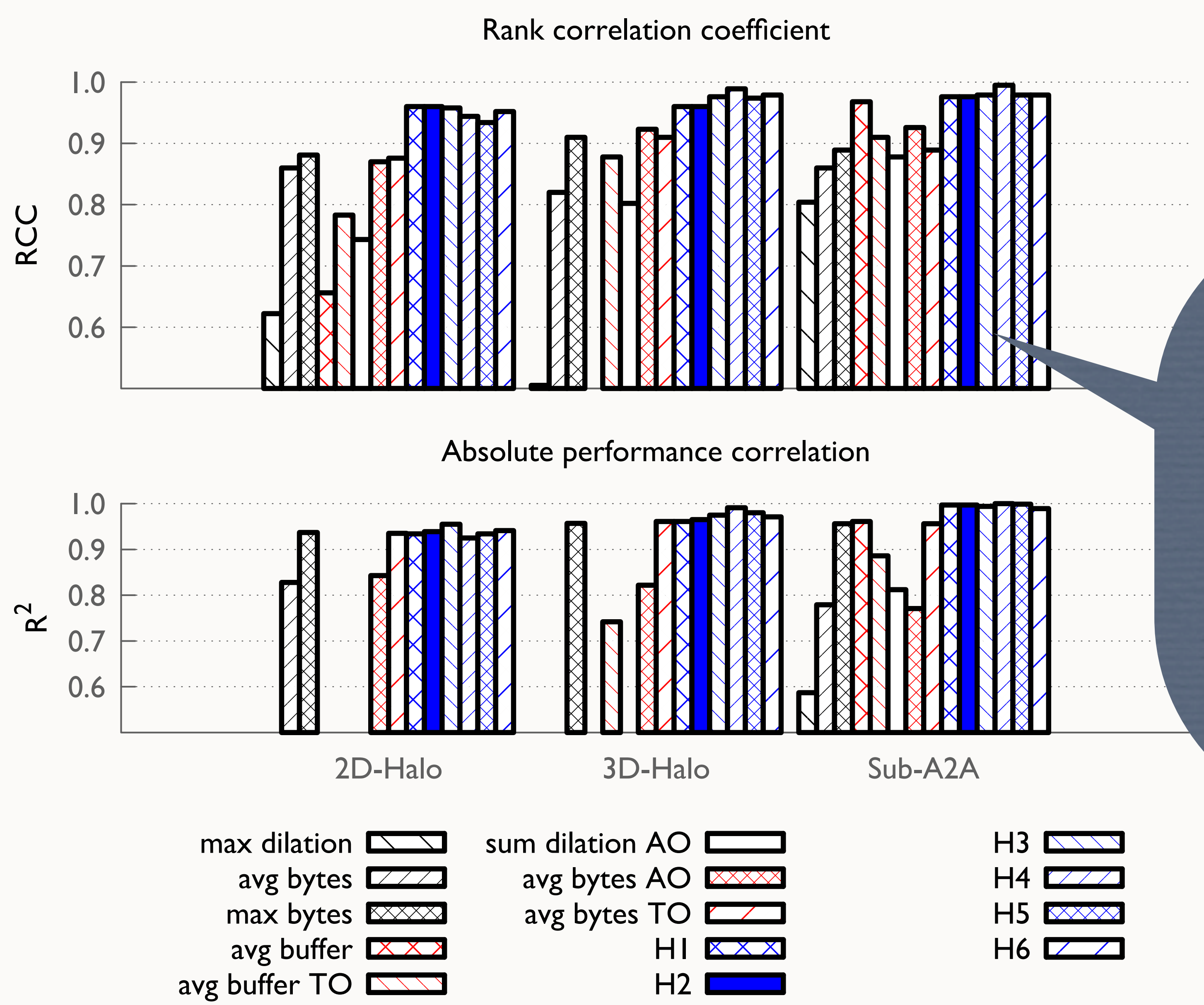
HYBRID METRICS

- Combine multiple metrics to complement each other
- Some combinations
  - avg bytes + max bytes + max FIFO
  - avg bytes + max bytes + avg buffer + max FIFO
  - avg bytes + avg buffer + avg delay AO + sum hops
  - avg bytes TO + avg buffer TO + avg delay TO + sum hops

RESULTS
HYBRID METRICS

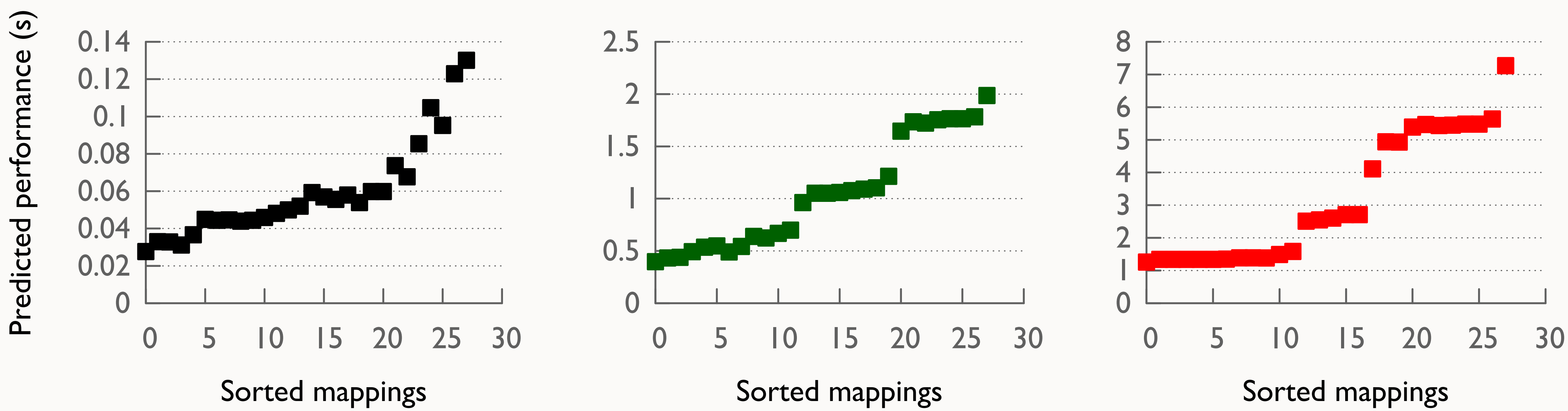


RESULTS
HYBRID METRICS



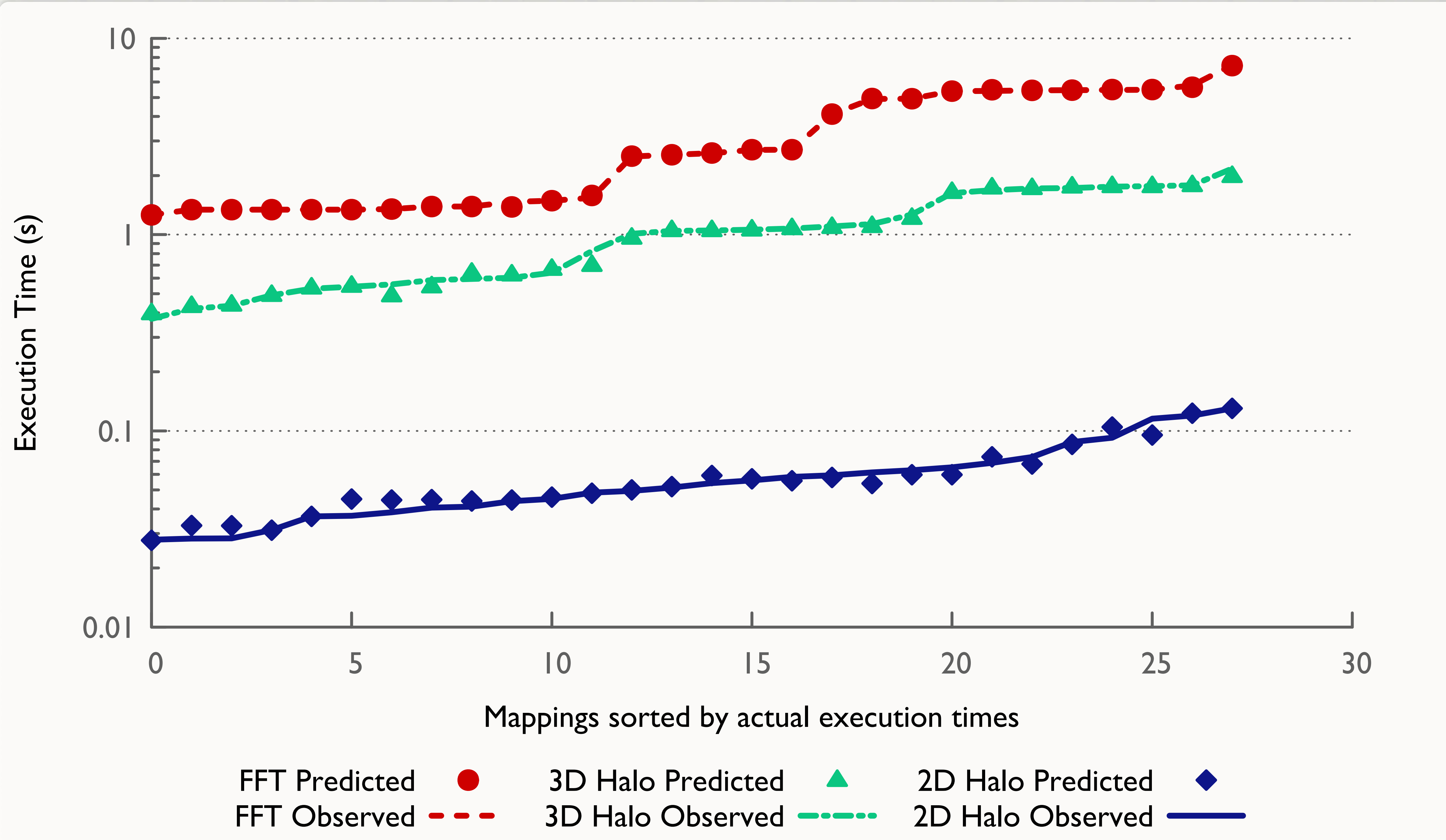
hybrid metrics provide high accuracy

RESULTS - TREND



2D Halo 3D Halo Sub A2A

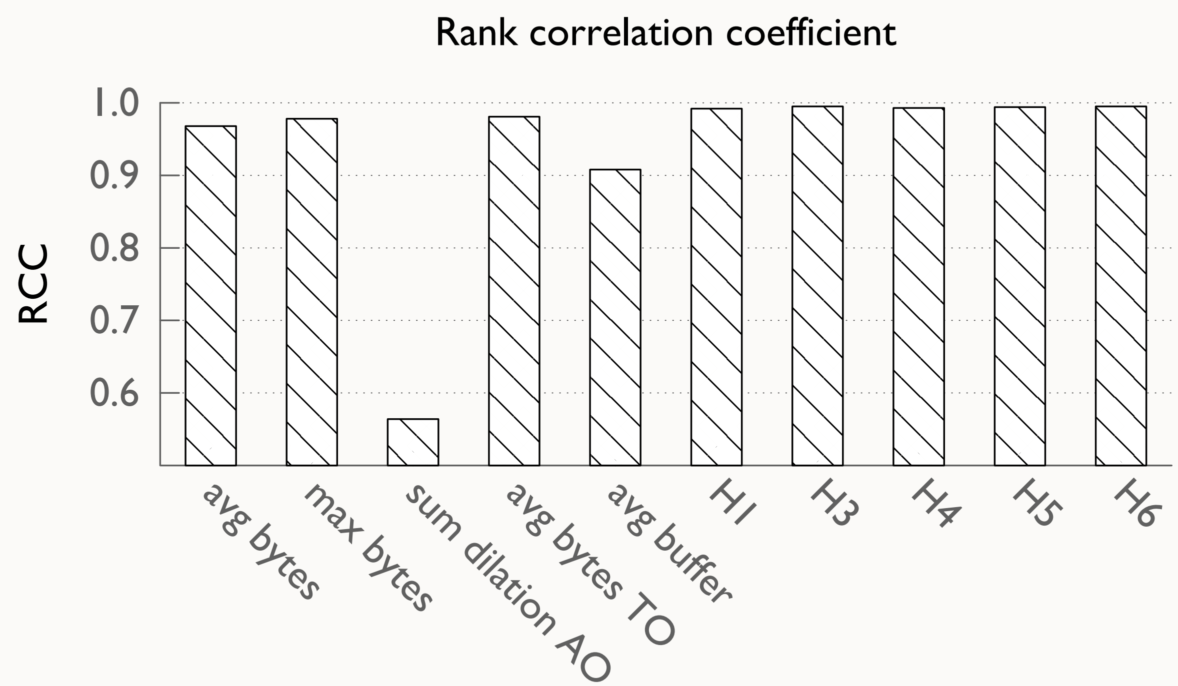
RESULTS



Execution Time (s)
10
1
0.1
0.01
0
5
10
15
20
25
30
Mappings sorted by actual execution times
FFT Predicted
FFT Observed
3D Halo Predicted
3D Halo Observed
2D Halo Predicted
2D Halo Observed

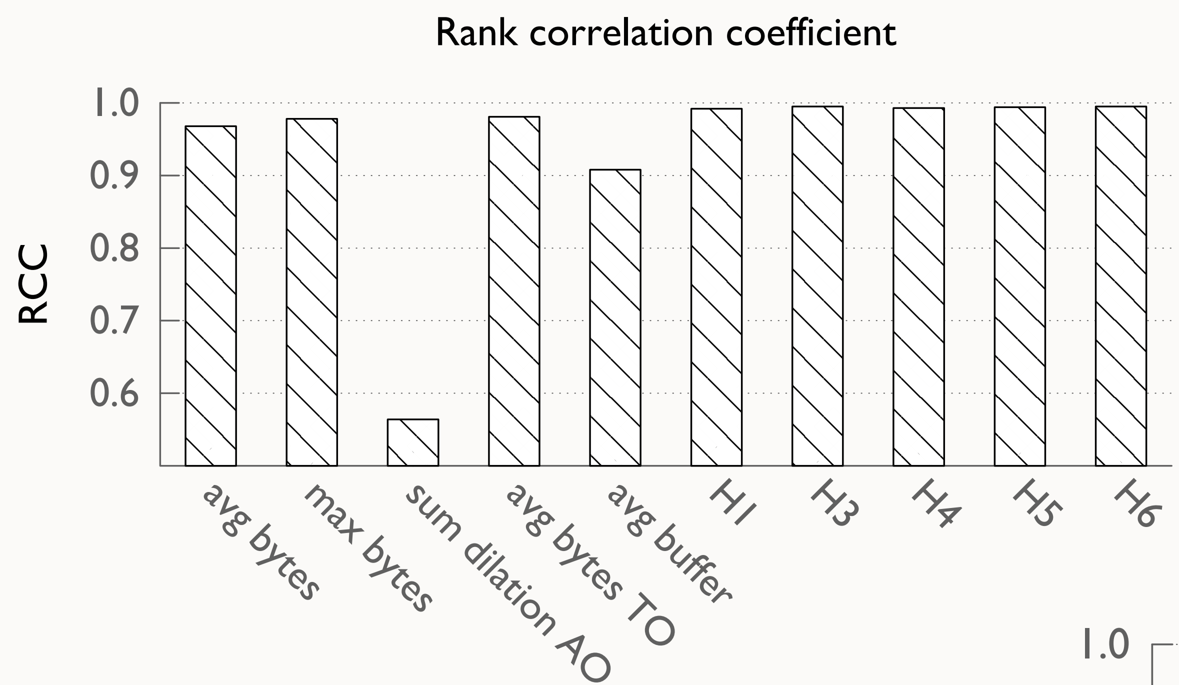
RESULTS

RESULTS

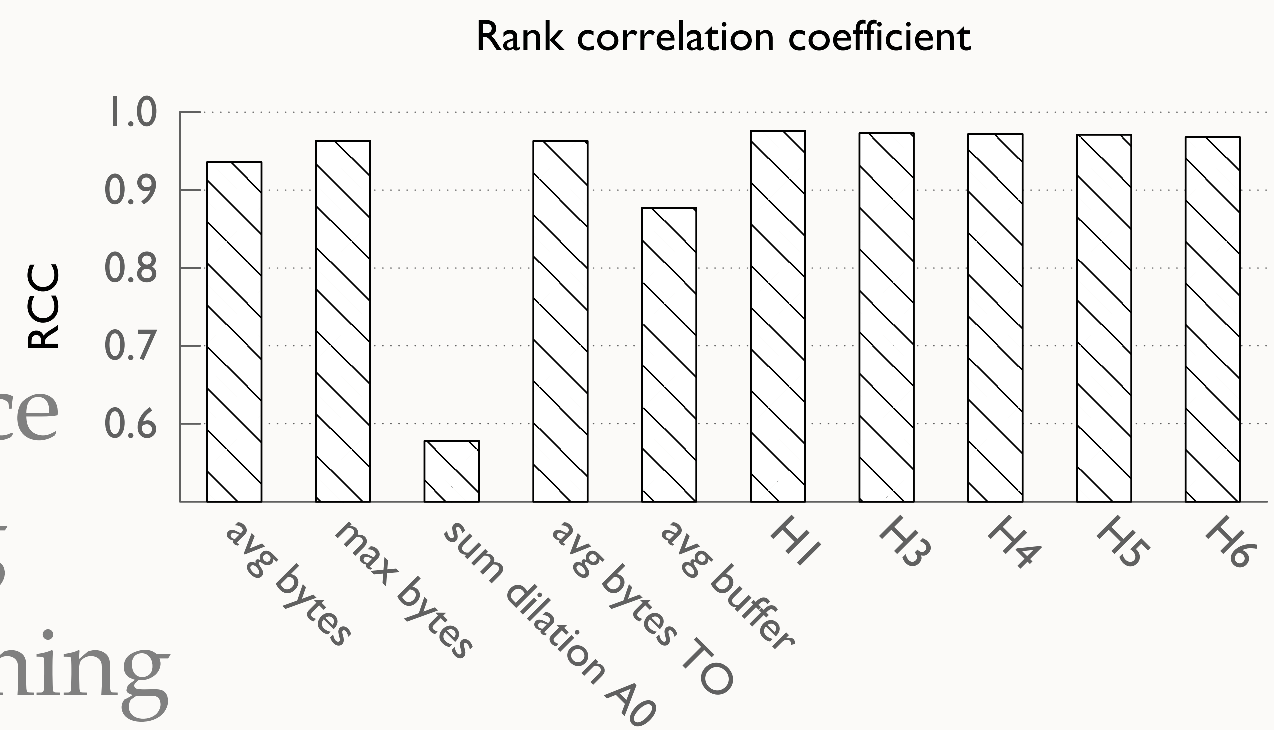


Combining all benchmarks

RESULTS



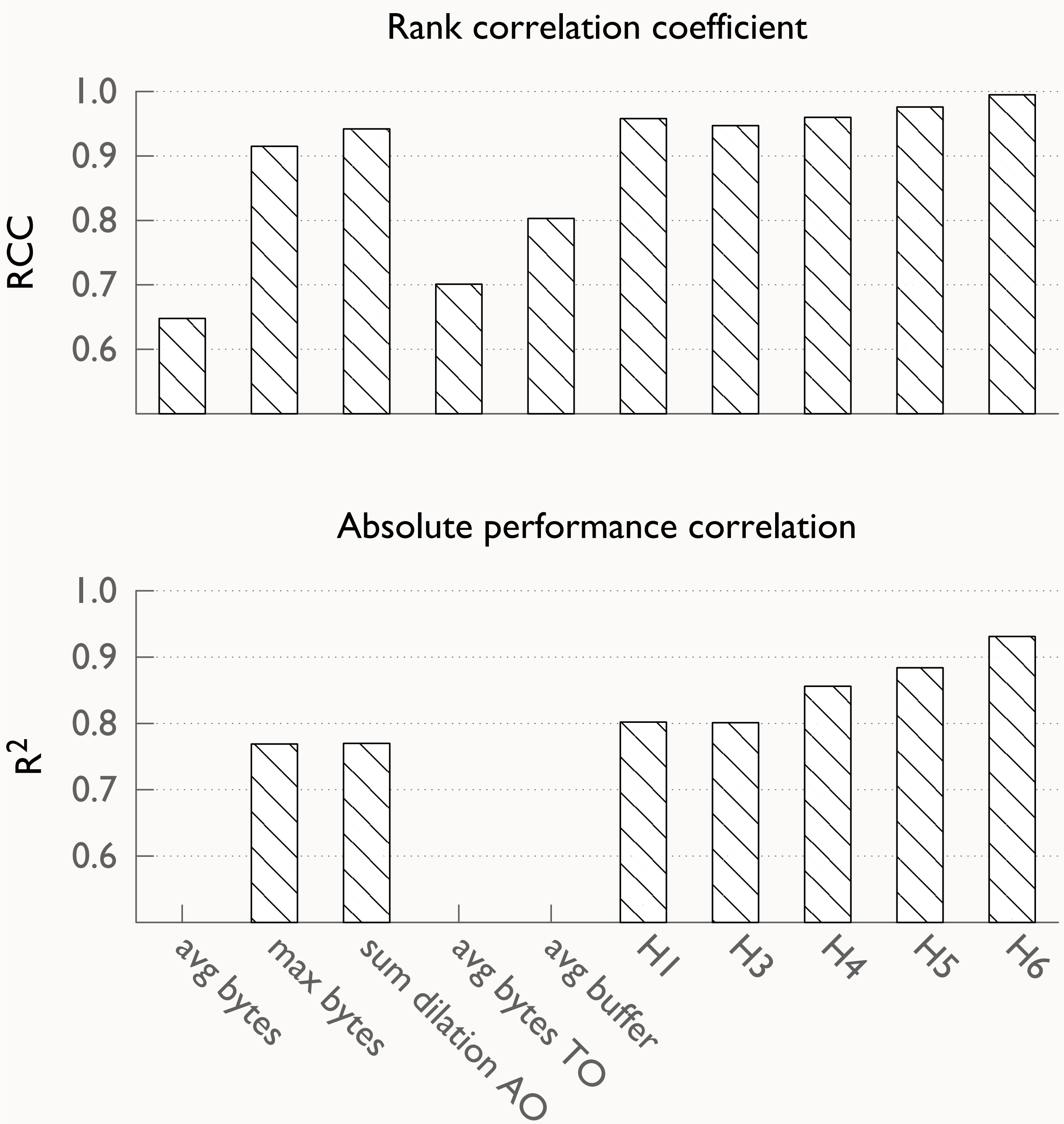
Combining all benchmarks



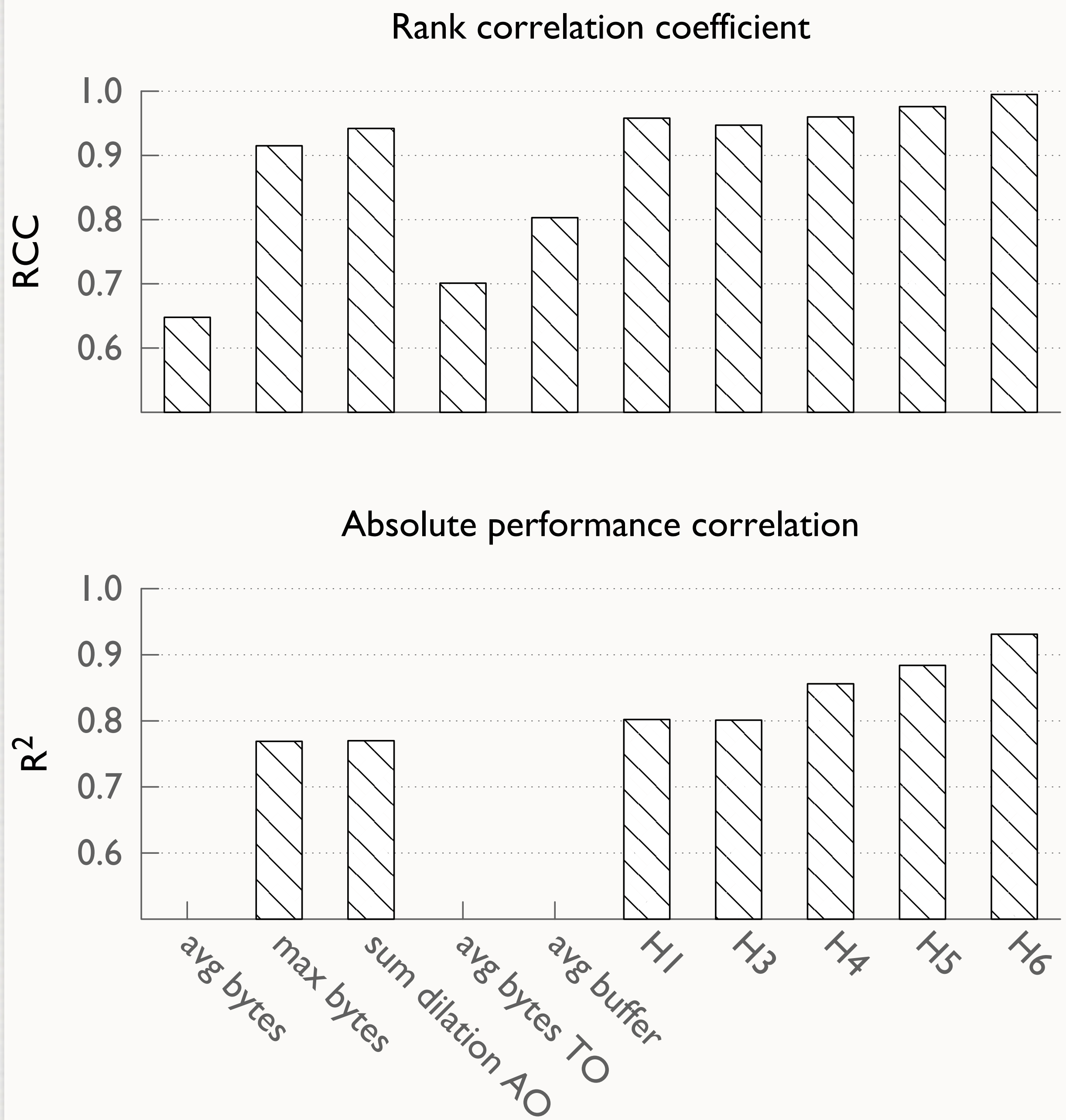
Predicting performance on 65,536 cores using 16,384 cores data for training

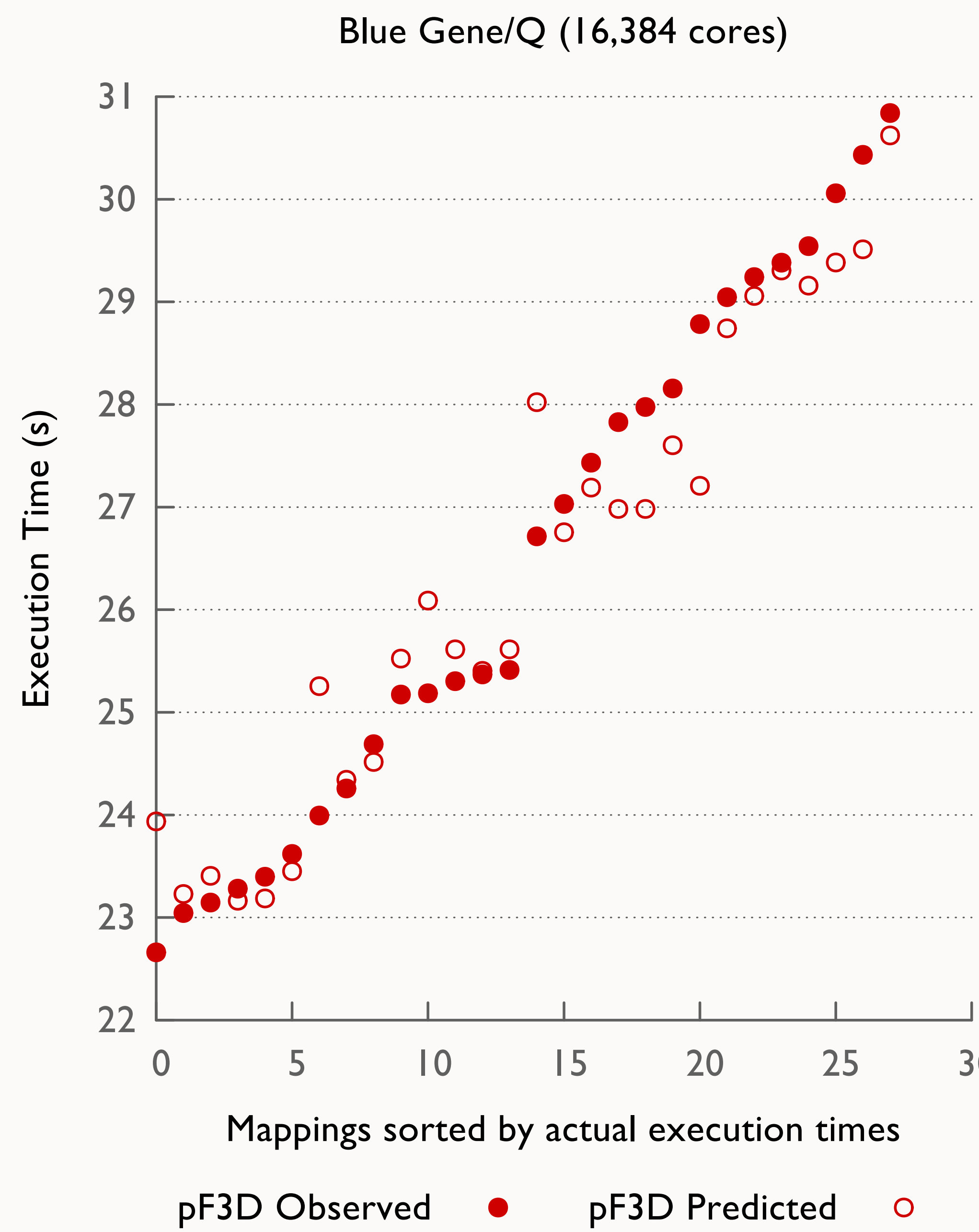
RESULTS - PF3D

RESULTS - PF3D



RESULTS - PF3D





SUMMARY

- Communication is not just about peak latency / bandwidth
- Simultaneous analysis of various aspects of network is important
- Complex models are required for accurate prediction
- There are patterns waiting to be identified!

FUTURE WORK

- More applications!
- More metrics
- Weighted analysis
- Offline prediction of entities

FUTURE WORK

- More applications!
- More metrics
- Weighted analysis
- Offline prediction of entities

Questions?