SONAR: Automated Communication Characterization for HPC Applications

Steffen Lammel, Felix Zahn, Holger Fröning

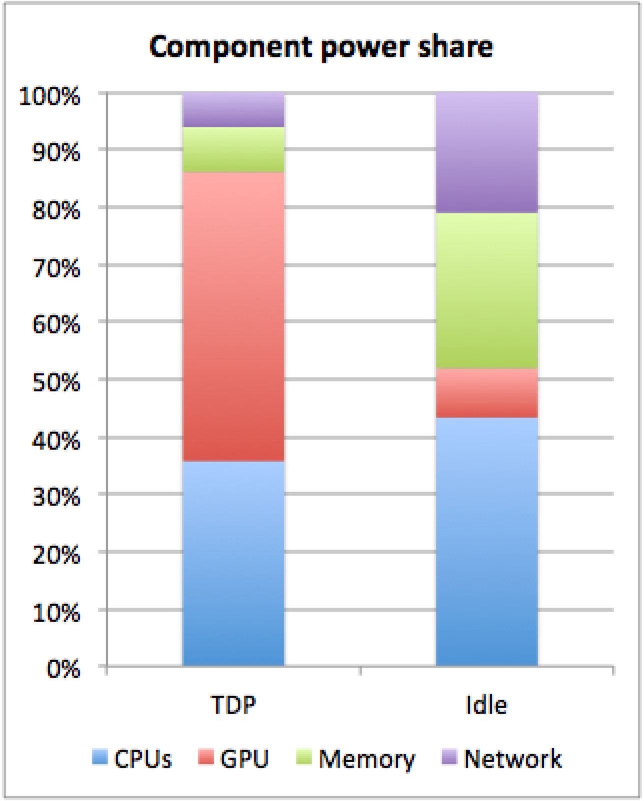
Computer Engineering Group, Ruprecht-Karls University of Heidelberg, Germany

ExaComm 2016 - Second International Workshop on Communication Architectures at Extreme Scale

06-23-2016

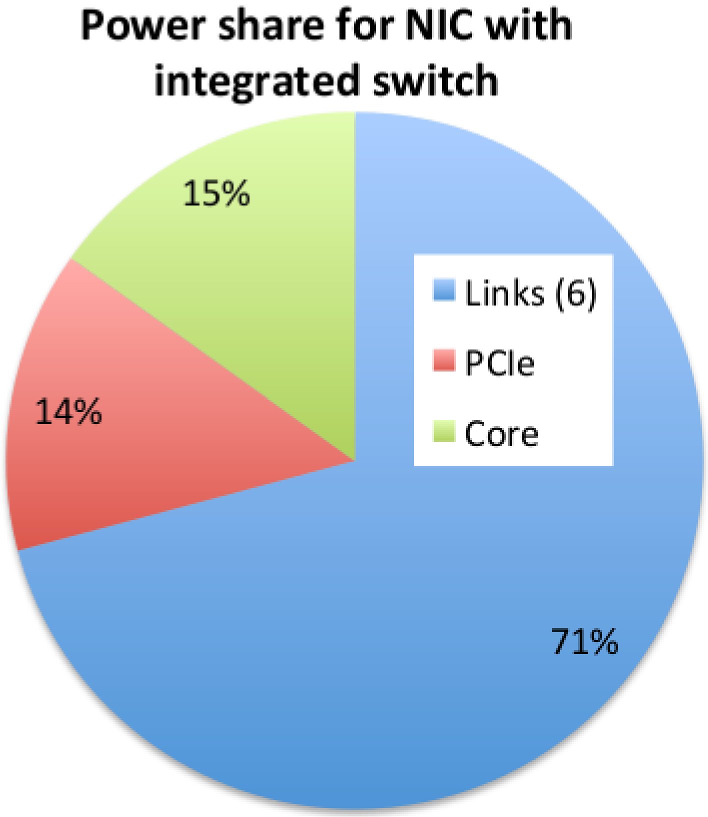
Motivation

- CPU power demands are well known today
- Interconnect is also important!
  - Up to 20-30% of the total power budget
  - Exascale is estimated to require 20-100MW



Saving Network Power

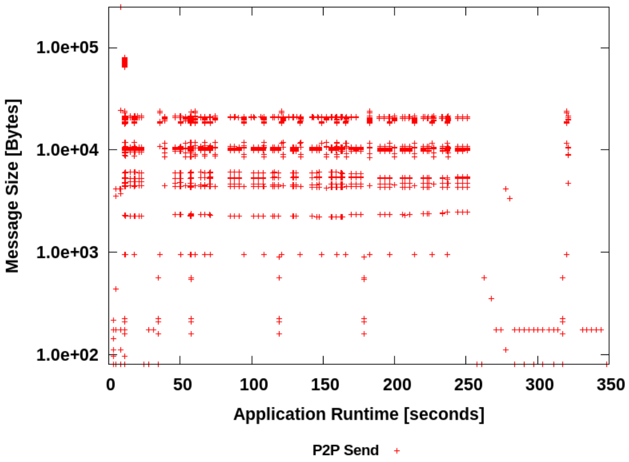
- Network power is driven by links
- Power consumption mainly depends on bandwidth
  - Link width
  - Link frequency

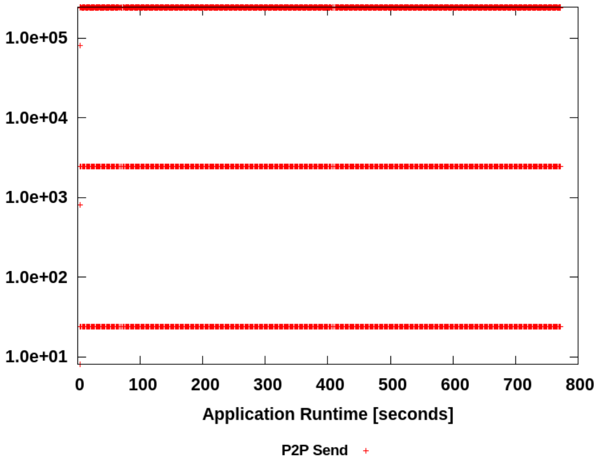


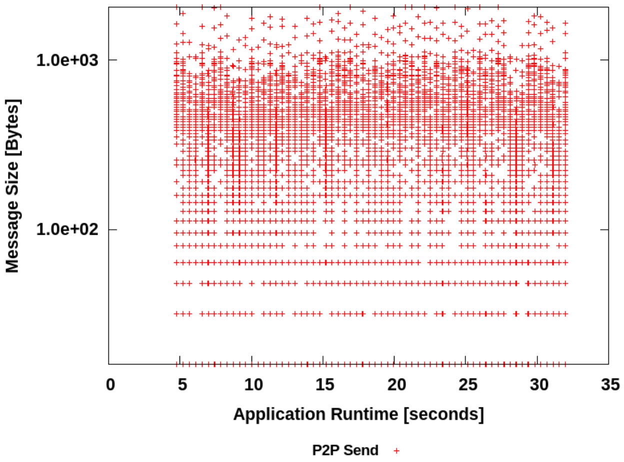
EXTOLL Tourmalet switch (TSMC 65nm process)

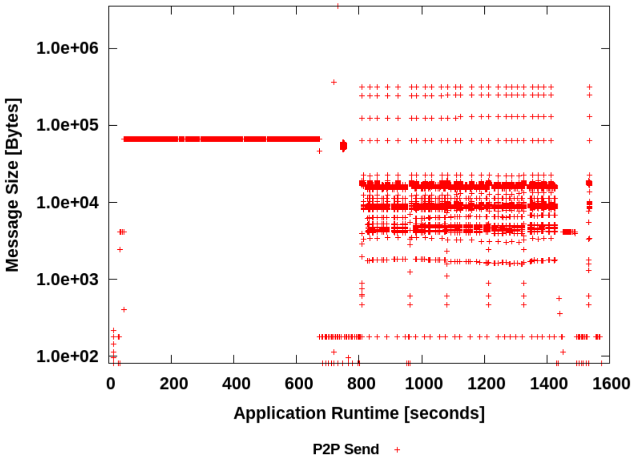
Zahn, F., Yebenes, P., Lammel, S., Garcia, P.J., Fröning, H.: Analyzing the energy (dis-) proportionality of scalable interconnection networks. In: 2nd IEEE International Workshop on High-Performance Interconnection Networks on the Exascale and Big-Data Era (HiPINEB). (2016)

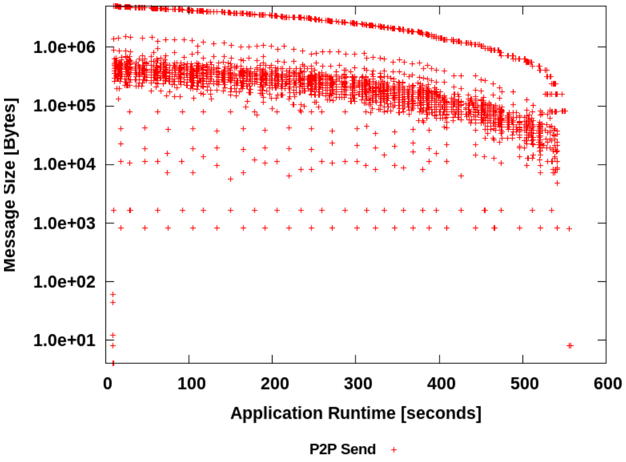
Traffic Patterns of Different Workloads

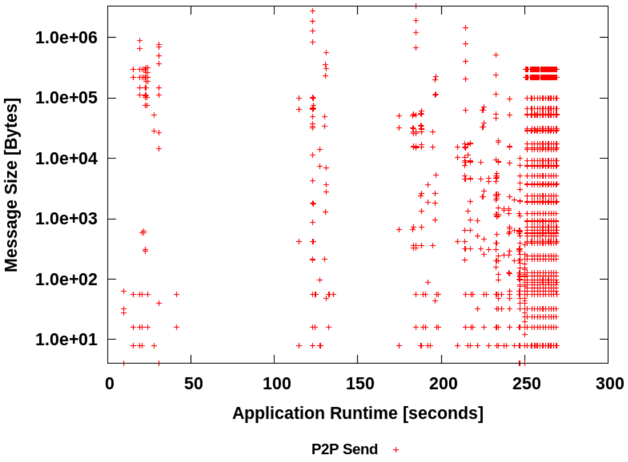






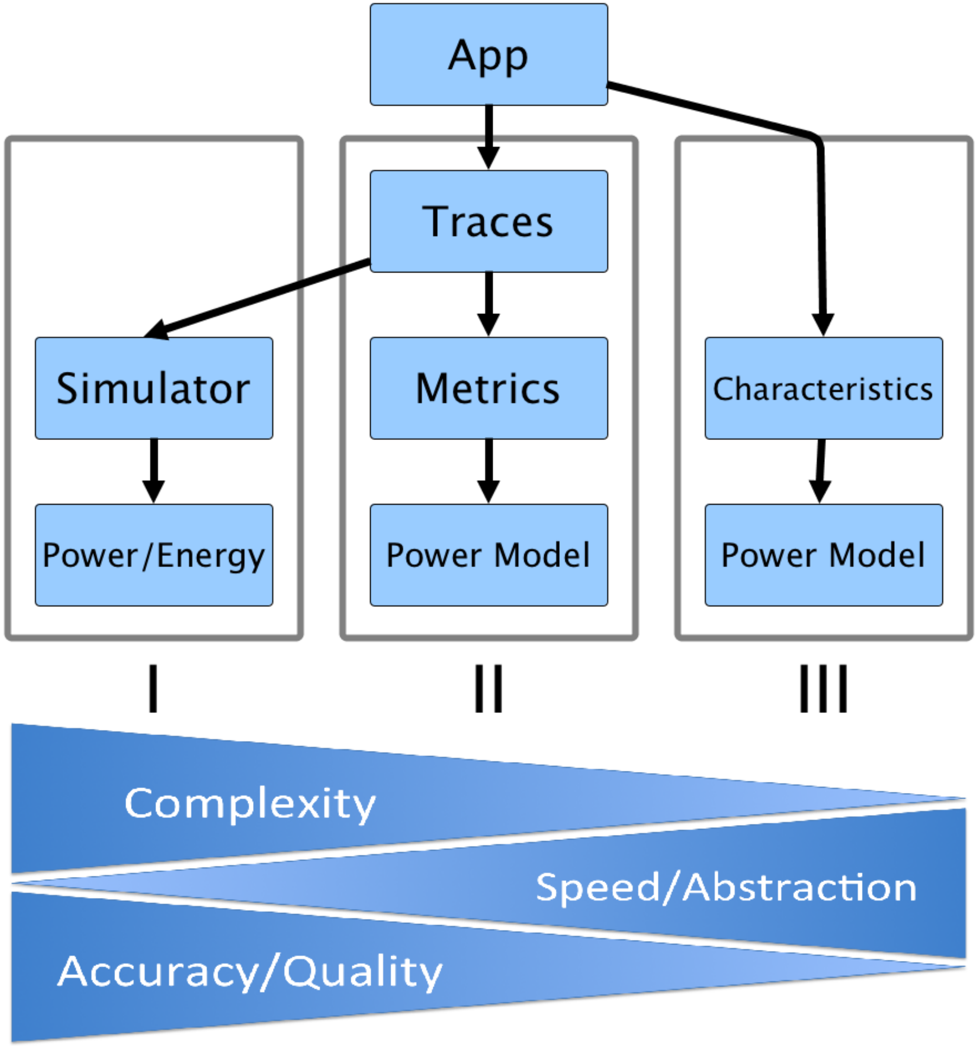






Methodology to Derive Power from Applications

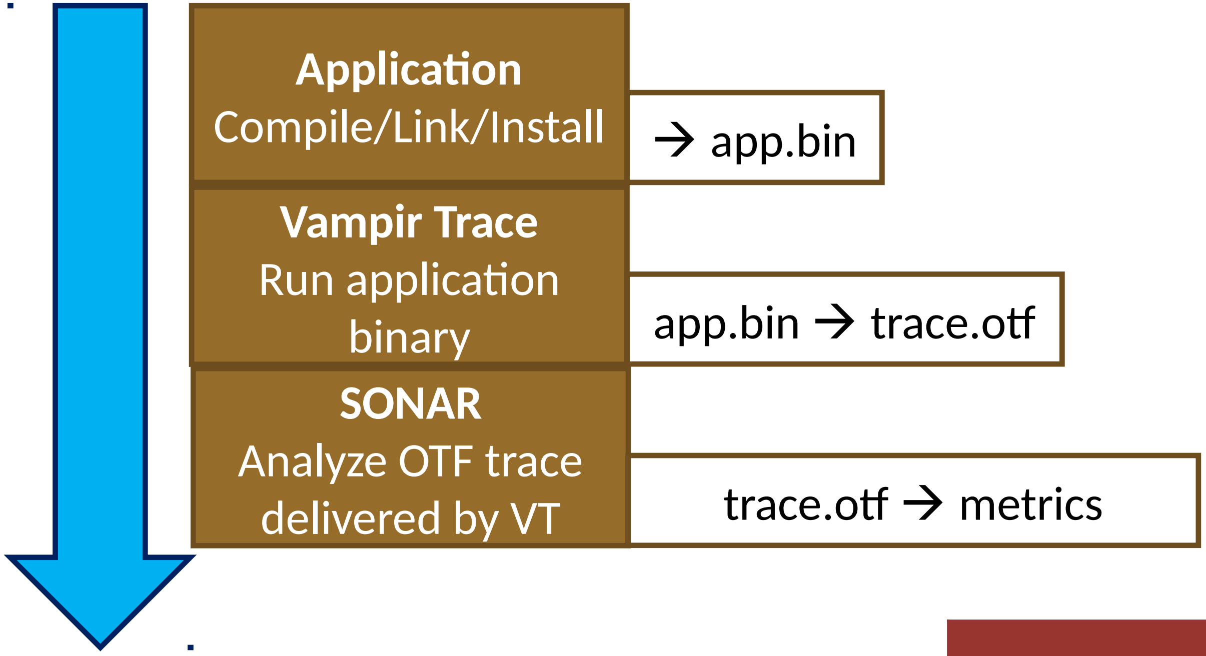
- Power-aware Simulator (I)
  - High accuracy, very long runtime
- Power Model based on Metrics (II)
  - Traces still necessary, first step to identify crucial characteristics
- Power Model (III)
  - Future goal
  - Allows deriving power consumption without running full application



Trace Generation

- Needed: tool that extracts custom metrics out of complex application behavior
  - We are not yet sure which metrics we might need
- Approach: modular tool based on parsing of application traces
  - Including communication and computation

SONAR workflow:



Application
Compile/Link/Install

→ app.bin

Vampir Trace
Run application binary

app.bin → trace.otf

SONAR
Analyze OTF trace delivered by VT

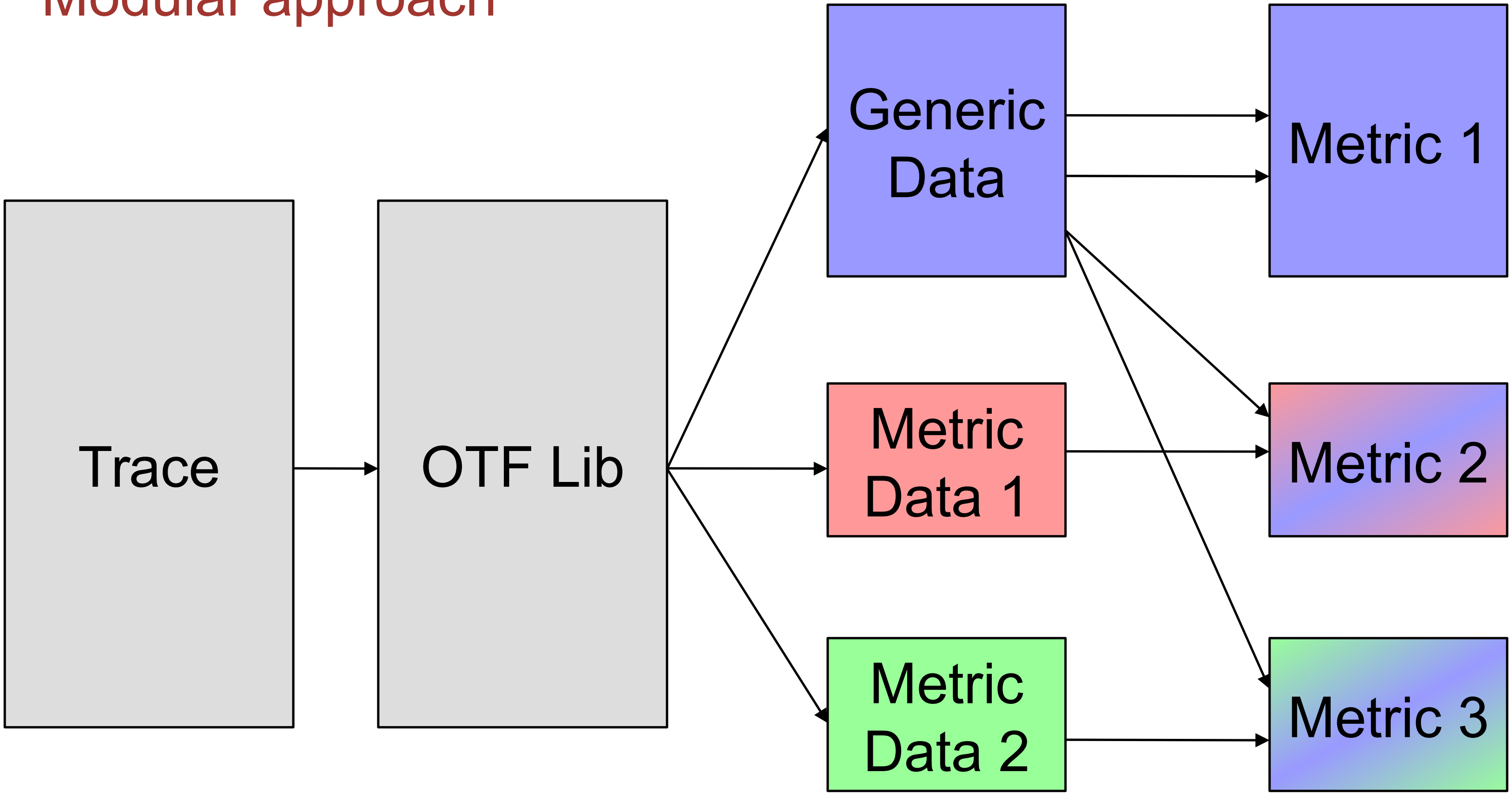
trace.otf → metrics

SONAR‘s Initial Set of Metrics

- Relevant for power model
  - Network Activity Map
  - MPI Idle Time
  - Application Verbosity (bytes/flop)
- Of general interest for the research group (rather “byproducts”)
  - Message Size Distribution
  - Message Rate

SONAR Overview

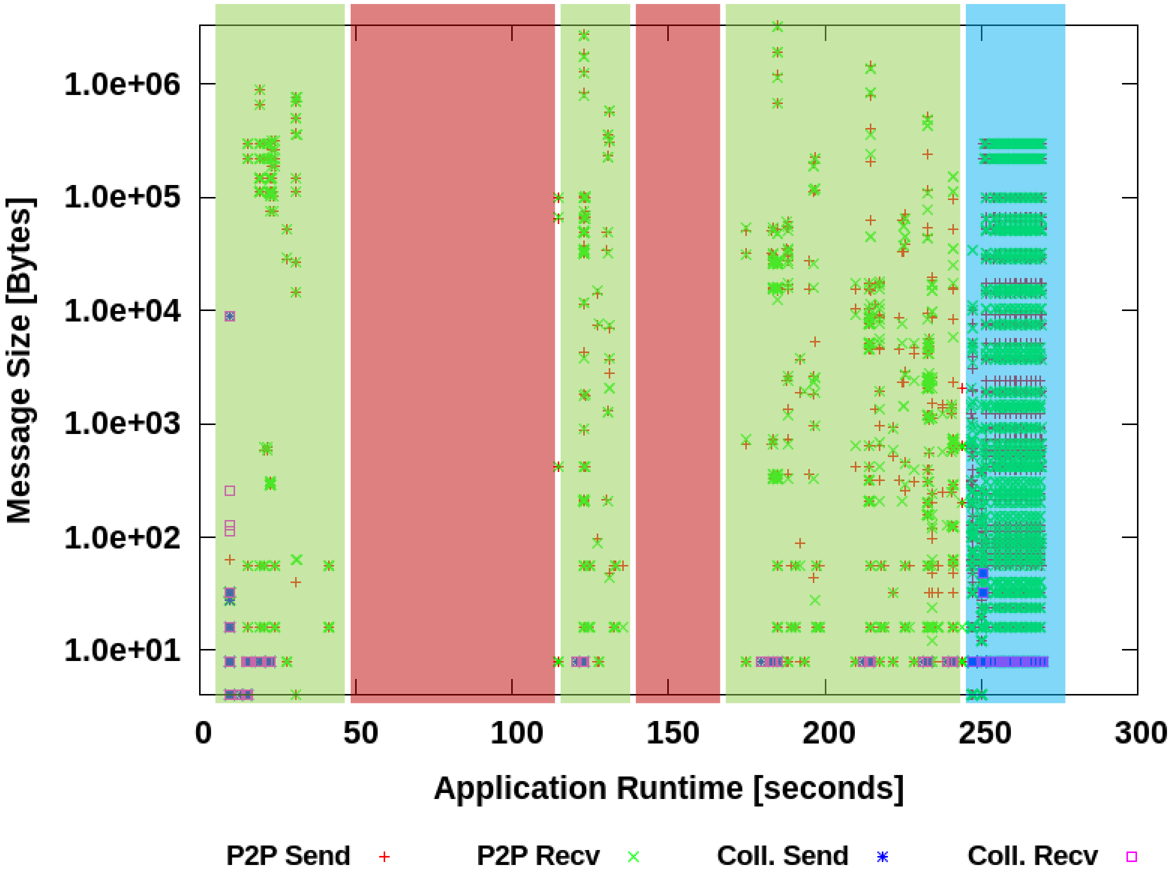
- Modular approach



Setup

- Benchmarks & Applications
  - High Performance Linpack (HPL)
  - Graph 500
  - NAMD (ApoA1 + STMV)
  - LULESH
  - AMG2013
- System
  - 8-node cluster
  - 2x Intel Xeon E5-2630v2, 64GB per node
  - Interconnect: GB-Ethernet

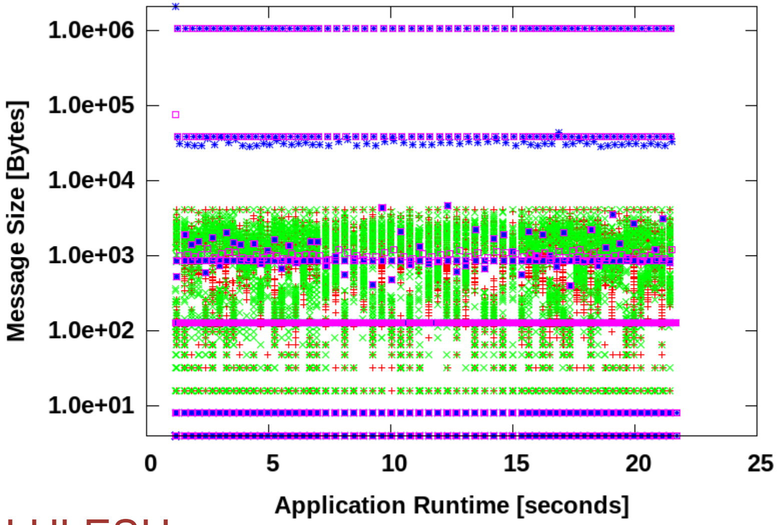
Metric: Network Activity Map



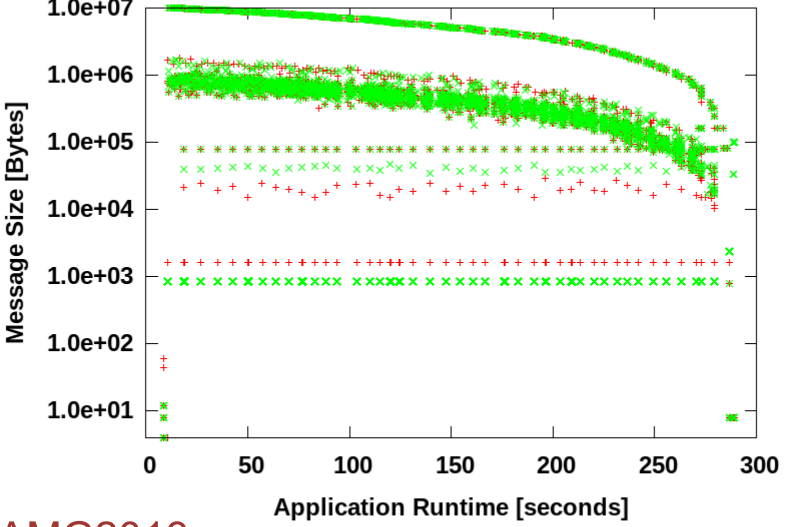
- Light Communication
- Dense Communication
- No Communication
  - MPI Idle Time

Exemplary Network Activity Maps

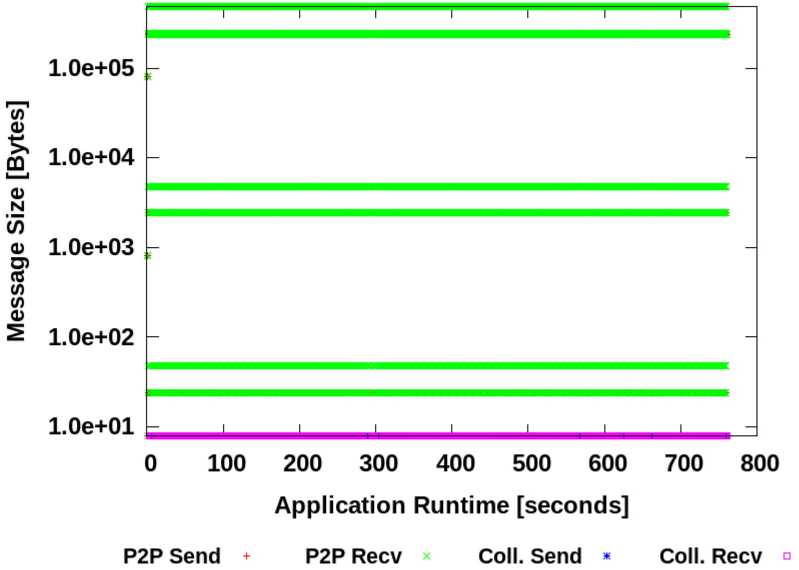
Graph 500



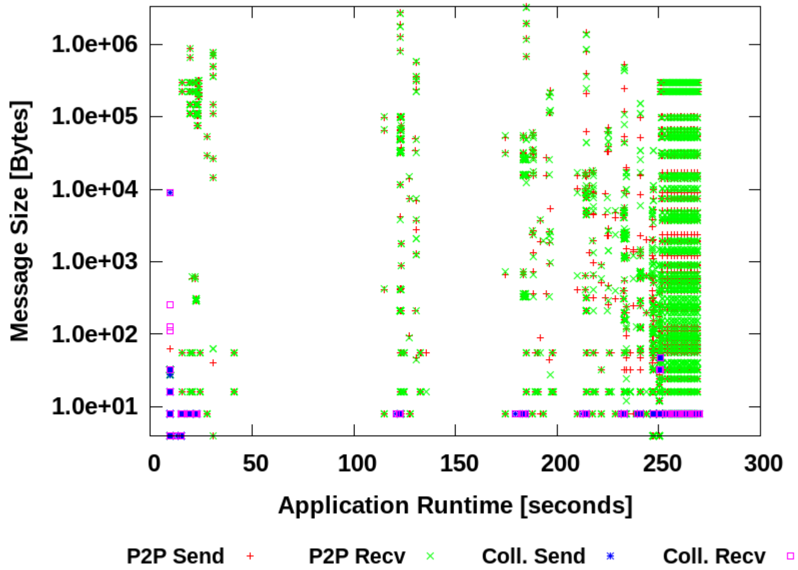
Linpack



LULESH

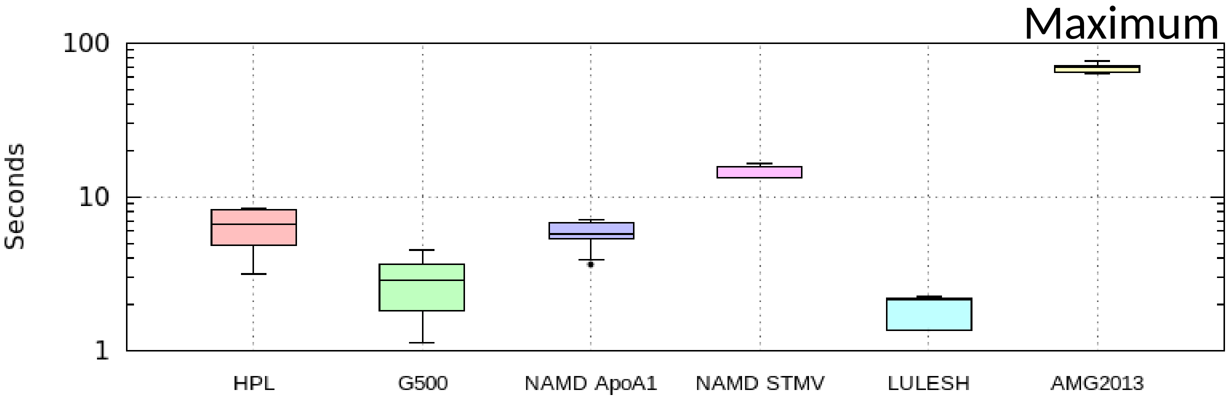


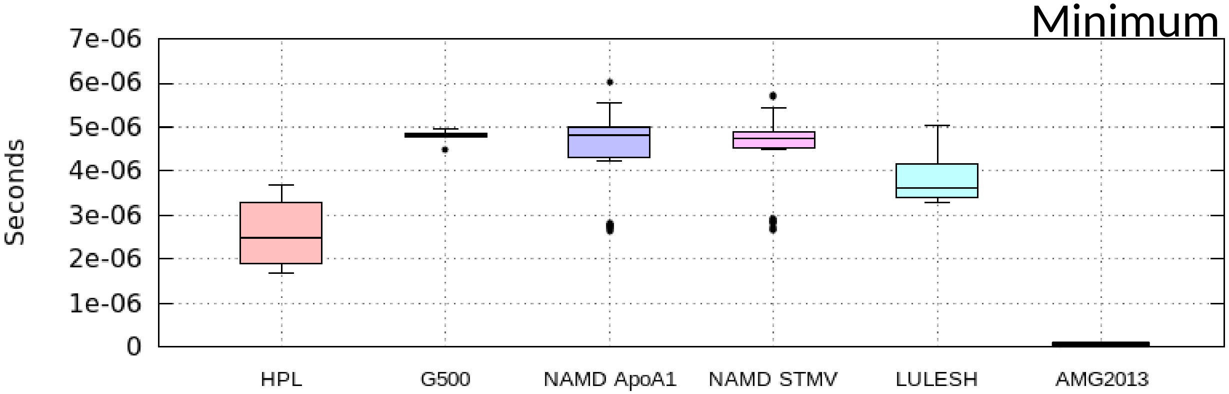
AMG2013

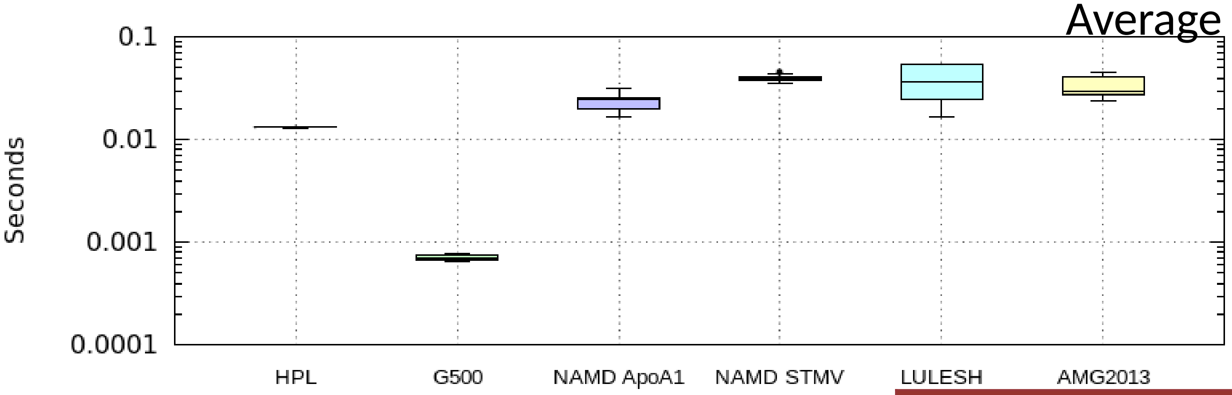


Metric: MPI Idle Time – Node Divergence

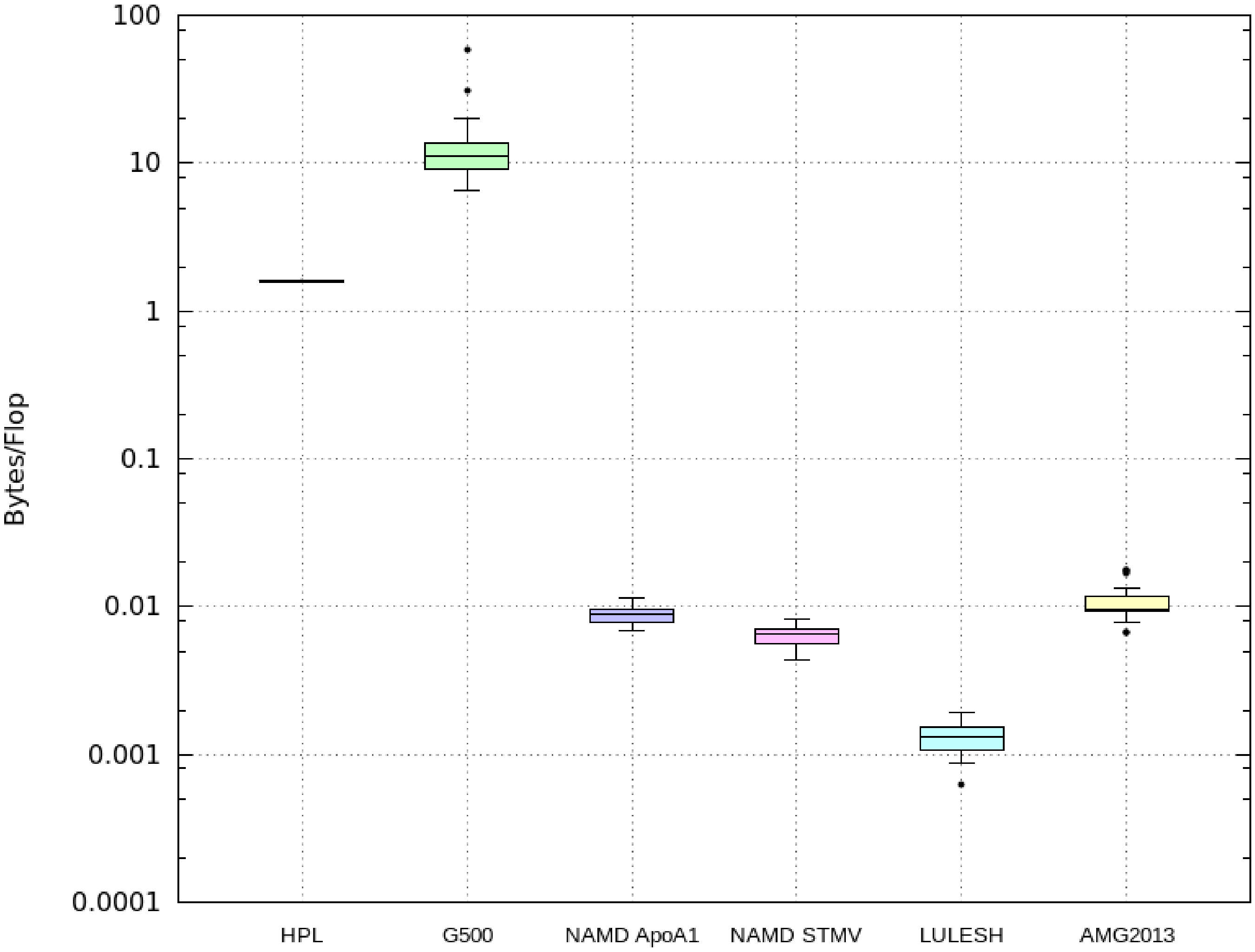
- Execution times:
  - HPL: 280s
  - Graph 500: 23s
  - NAMD ApoA1: 90s
  - NAMD STMV: 370s
  - LULESH: 780s
  - AMG2013: 270s


Maximum


Minimum

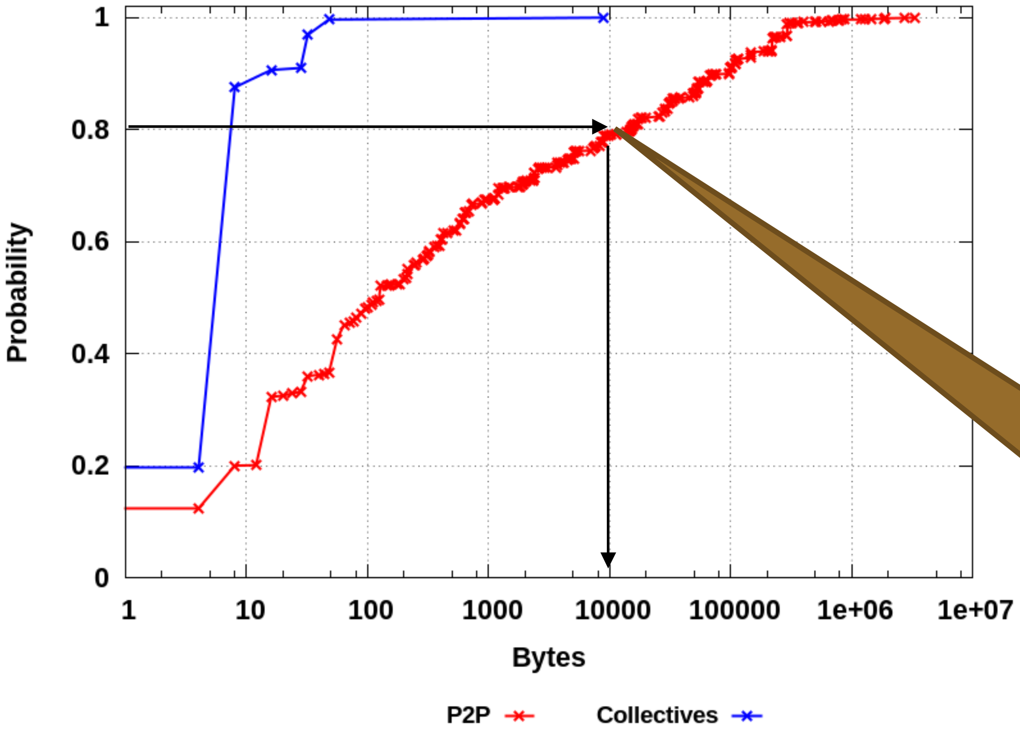

Average

Metric: Verbosity – Node Divergence



Metric: Message Size Distribution

- Message size distribution as CDF graph
  - Percentage of messages which are of size X or smaller

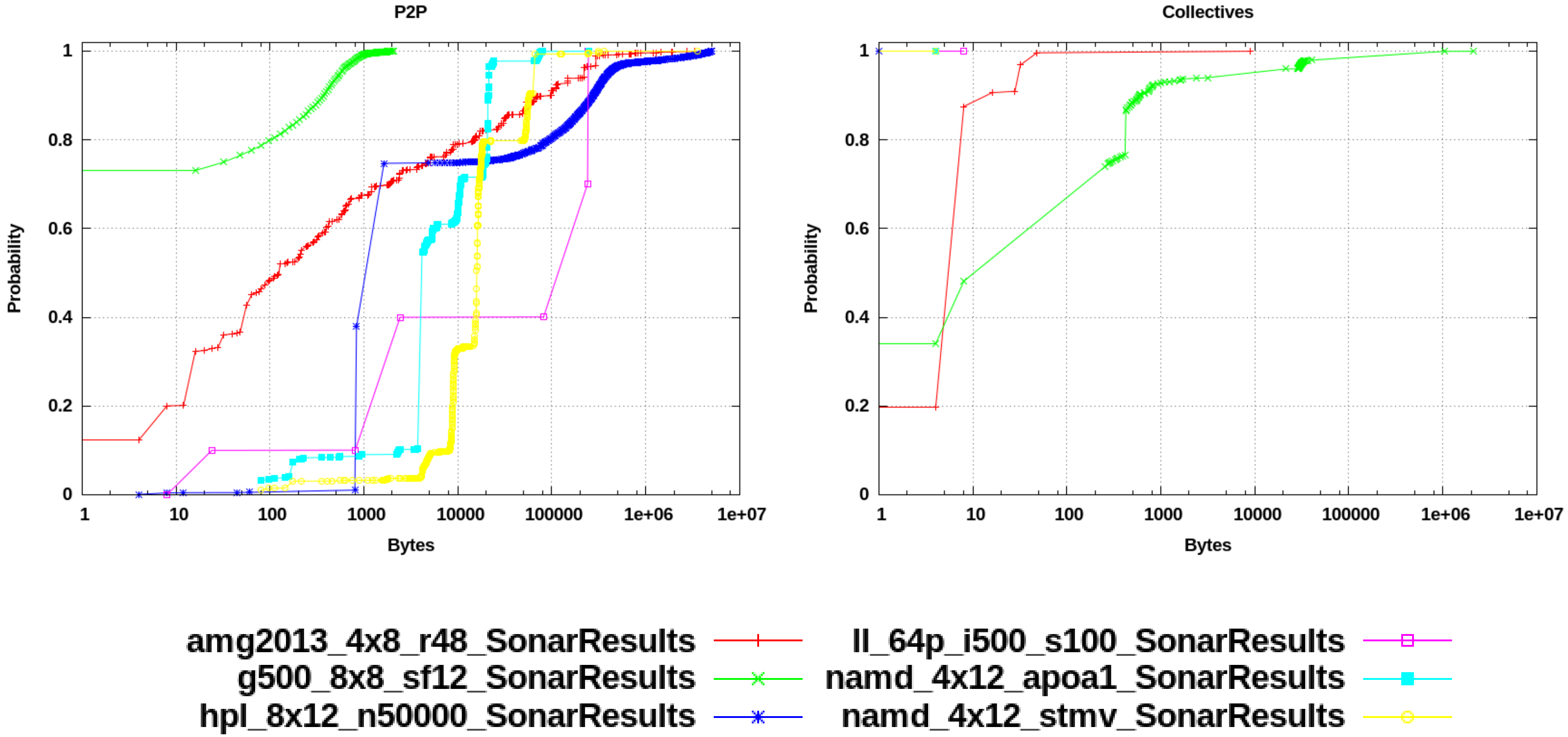


e.g. 80% of the P2P messages are smaller than 10k Bytes

Metric: Message Size Distribution

Point-to-Point

Collective



Conclusion

- MPI traces show very good potential for power saving in networks
  - Long MPI idle times
  - Strong correlation among nodes
- SONAR
  - Analyzes complex communication characteristics of HPC applications
  - Supports easy integration of new metrics
  - Is a first step to a power-aware network model
- Outlook
  - Understanding the impact of current metrics to the network power consumption
  - Further exploration of suitable metrics

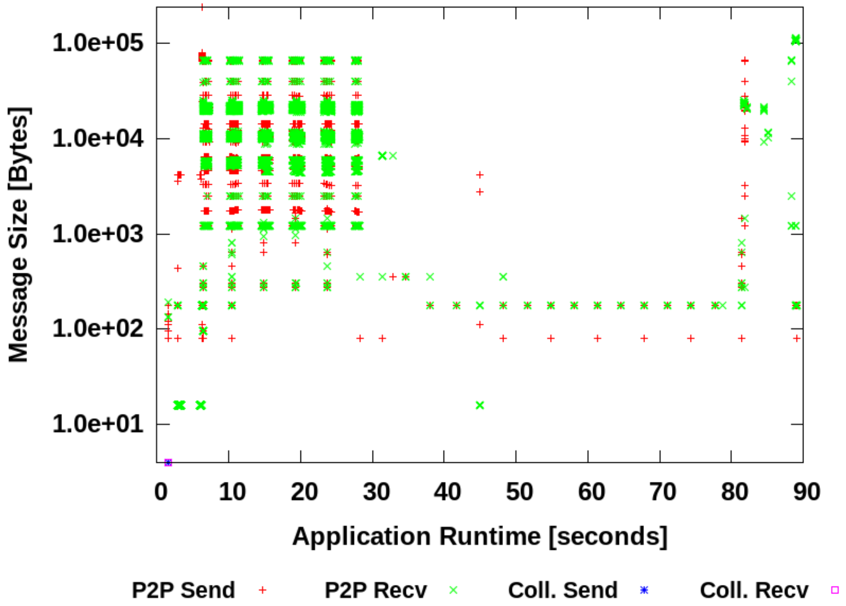
Thank you for your attention!

Questions?

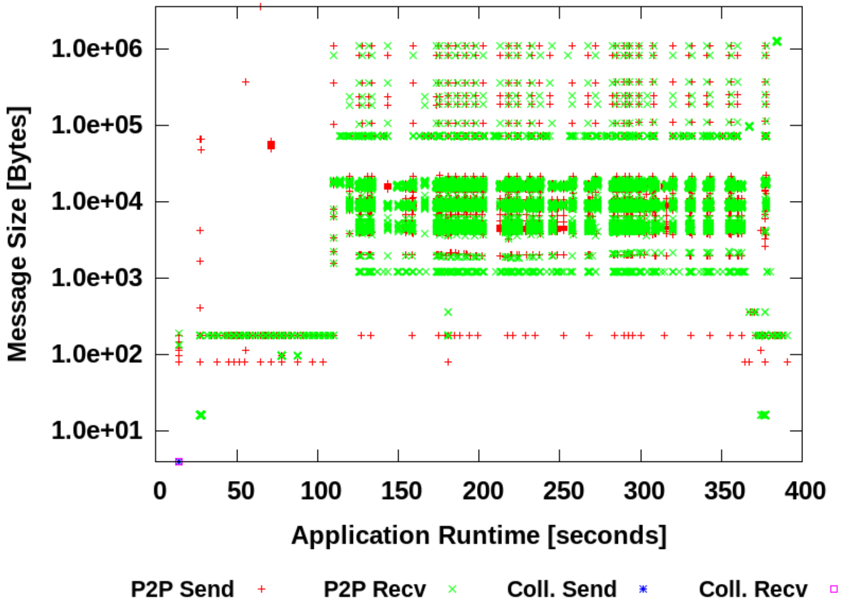
Spare Slides

Network Activity Map

ApoA1

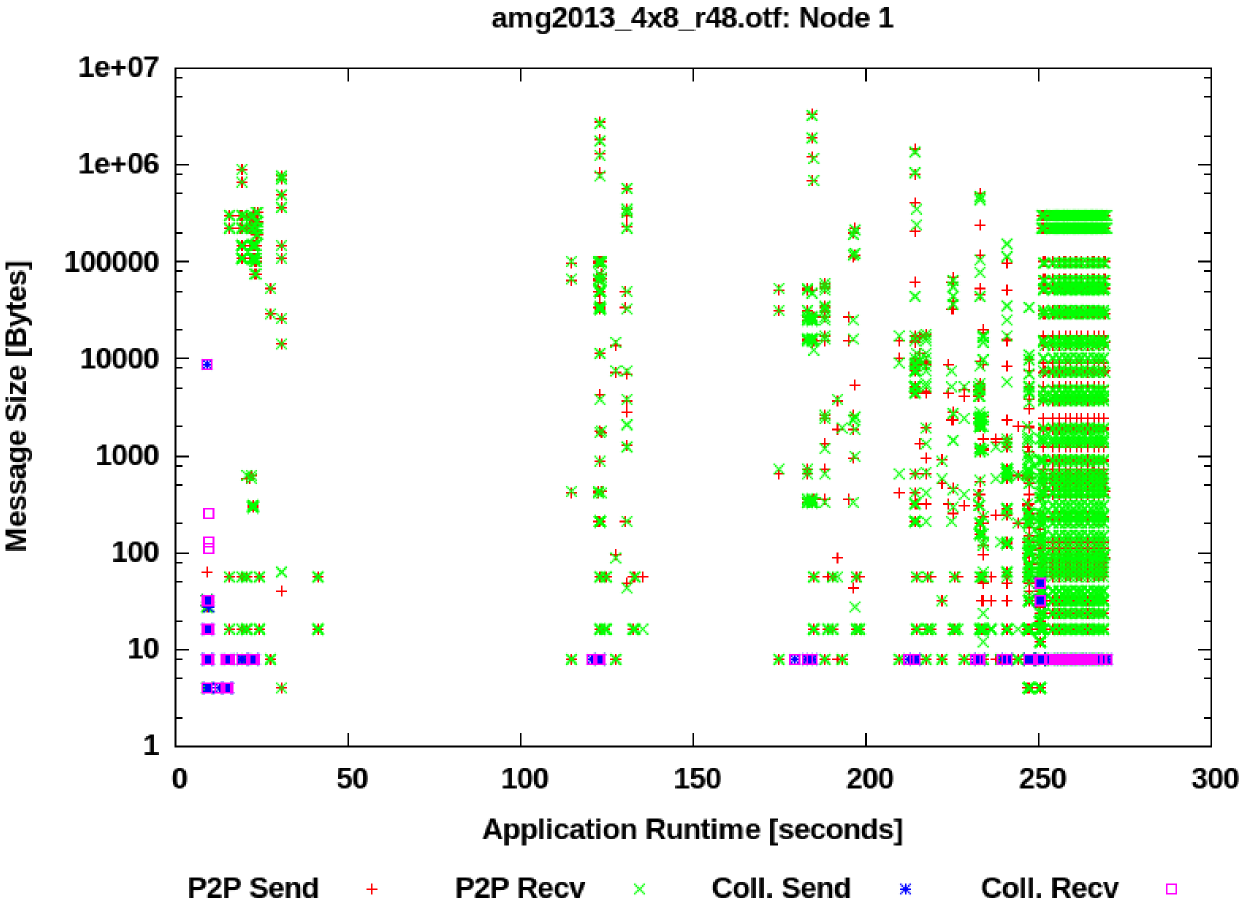


STMV



- Same application (NAMD), but with different input data

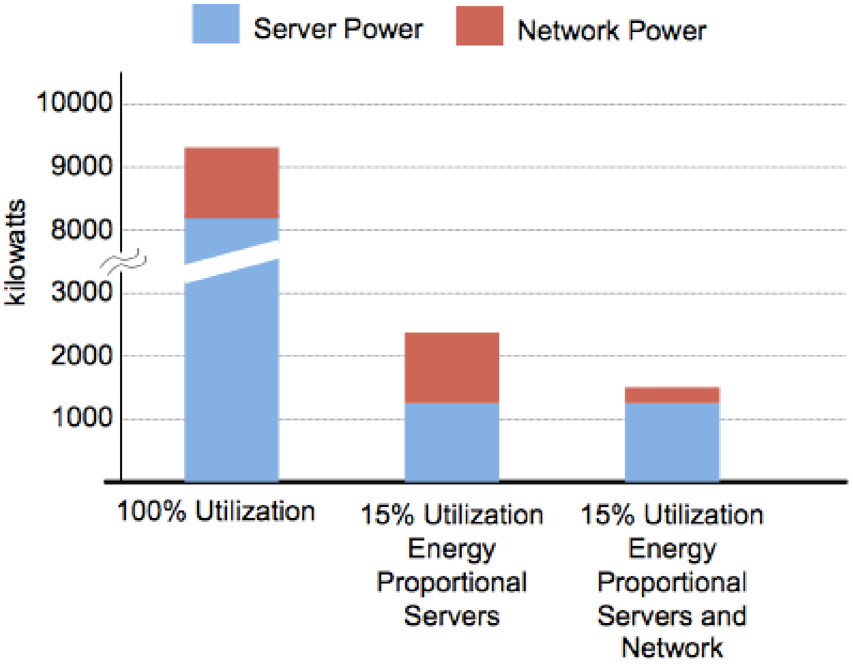
Network Activity Map



Seconding opinion on System Power

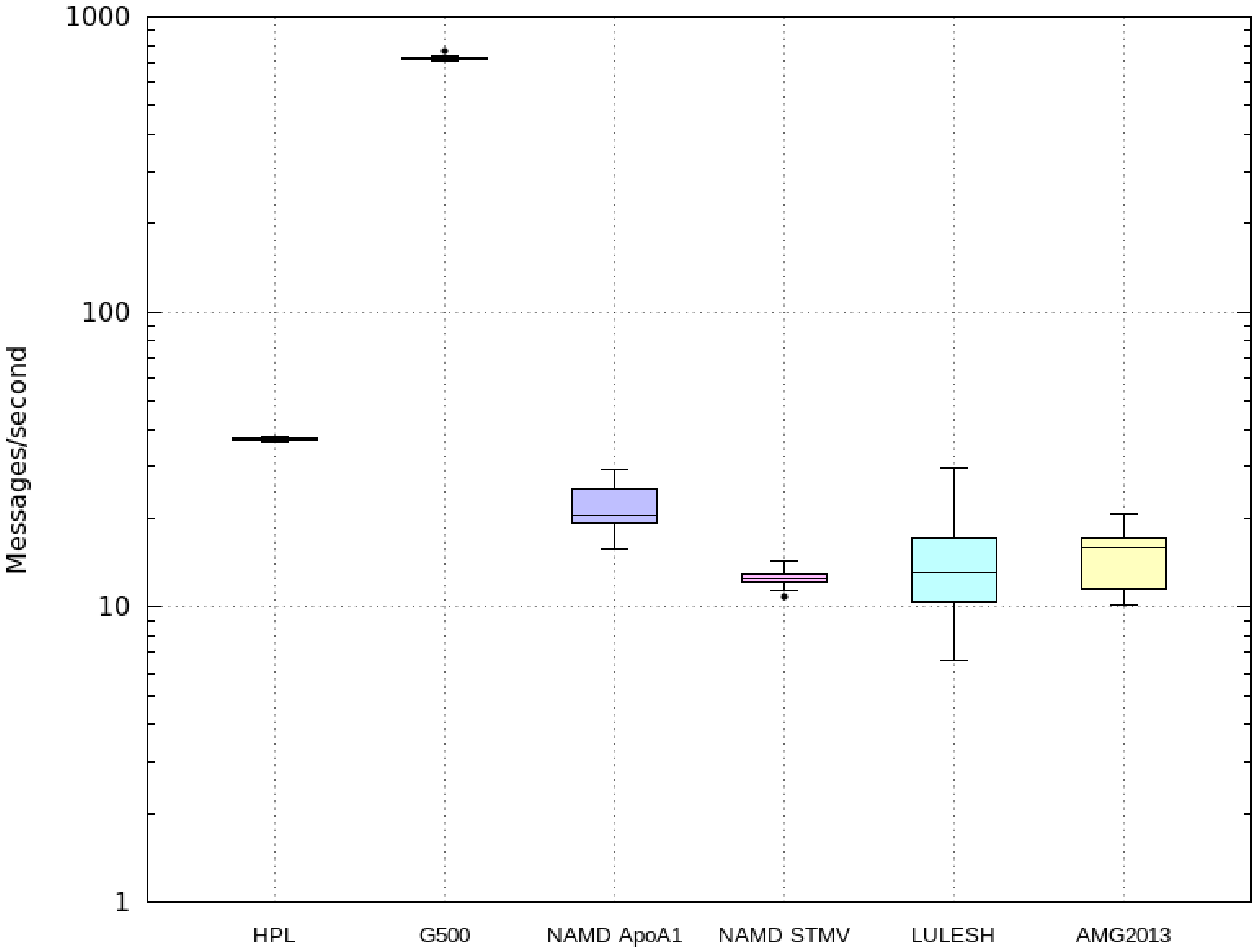
- We need energy-proportional components
  - Processors have already improved significantly
- Lesson learned from embedded systems: anything matters

- **Google paper on energy-proportional networks**: up to 50% on network power,32k nodes: 1.1MW for folded CLOS, 0.7MW for flattened butterfly
- **S. Rumsey et. al. (ISC2015)**: networks continue to consume ~20% of system power even using optical links
- **DOE Report on Top 10 Exascale Challenges**: “Interconnect technology: Increasing the performance and energy efficiency of data movement”



Dennis Abts, Michael R. Marty, Philip M. Wells, Peter Klausler, and Hong Liu. 2010. Energy proportional datacenter networks. ISCA ’10
S. Rumley. et al., "Design Methodology for Optimizing Optical Interconnection Networks in High Performance Systems”, ISC 2015.

Metric: Message Rate


Messages/second
HPL
G500
NAMD ApoA1
NAMD STMV
LULESH
AMG2013