

Performance Analysis of MPI Collective Operations

Jelena Pješivac-Grbović, Thara Angskun, George Bosilca, Graham Fagg, Edgar Gabriel, Jack Dongarra



INNOVATIVE COMPUTING LABORATORY

COMPUTER SCIENCE DEPARTMENT
UNIVERSITY OF TENNESSEE

- » Motivation
- » Approaching the problem
- » Background and relevant work
- » Our research
 - » Optimized collective communication library
 - » Collective algorithm modeling
- » Results
- » Summary
- » Future work

- » Important component of MPI library
 - » More than 90% of MPI jobs use them
 - » If not optimized, can become performance bottleneck
- » It is an interesting problem

Optimal implementation of a given collective operation depends on many factors:

- » Underlying network topology
- » Number of involved processors
- » Data (size and format)
- » Application communication pattern
- » Load balance among the nodes
- » Network traffic

Approach to the Problem?

**Communication
Models**

**Computation
Models**

**Collective
Operation
Model**

Keep in mind model limitations!!!

» Hockney

$$T = \alpha + \beta \cdot m$$

$\alpha(m)$ – start up time

$\beta(m)$ – 1/bandwidth

» PLogP

$$T = L + g(m)$$

L – end to end latency

$g(m)$ – gap per message

$os(m)$, $or(m)$ – sender and receiver overheads

» LogP

$$T = L + 2 \cdot o$$

» LogGP

$$T = L + 2 \cdot o + G \cdot (m - 1)$$

L – latency

o – overhead

g – gap per message

G – gap per byte

» Linear model

$$T = \gamma \cdot m_s$$

- computation time per byte

» Complex computation models

» Operation

» Data type

» Application access pattern

» Cache effects

- » Thakur, et al. Hockney model
- » Chan, et al. Hockney model
- » Cameron, et al. Log_nP
- » Vadhiyar, et al. LogP extensions
- » Bernaschi, et al. LogGP
- » Kielmann et al. PLogP model

- » Optimized Collective Communication Library
- » Built on top of MPI point-to-point operations
- » Modularized
 - » Methods, Verification, and Performance
- » Currently supported collectives
 - » MPI Barrier, MPI Bcast, MPI Reduce, MPI_Scatter, MPI_Alltoall, and MPI_Allreduce

» Barrier

- » Linear, Ring, Recursive Doubling, Bruck

» Alltoall

- » Linear, Ring, Pairwise Exchange, Index (Bruck)

» Scatter

- » Linear

» Broadcast

- » Linear, Binomial, Binary, k-Chain (Pipeline)

» Reduce

- » Linear, Binomial, Binary, k-Chain (Pipeline)

» Allreduce

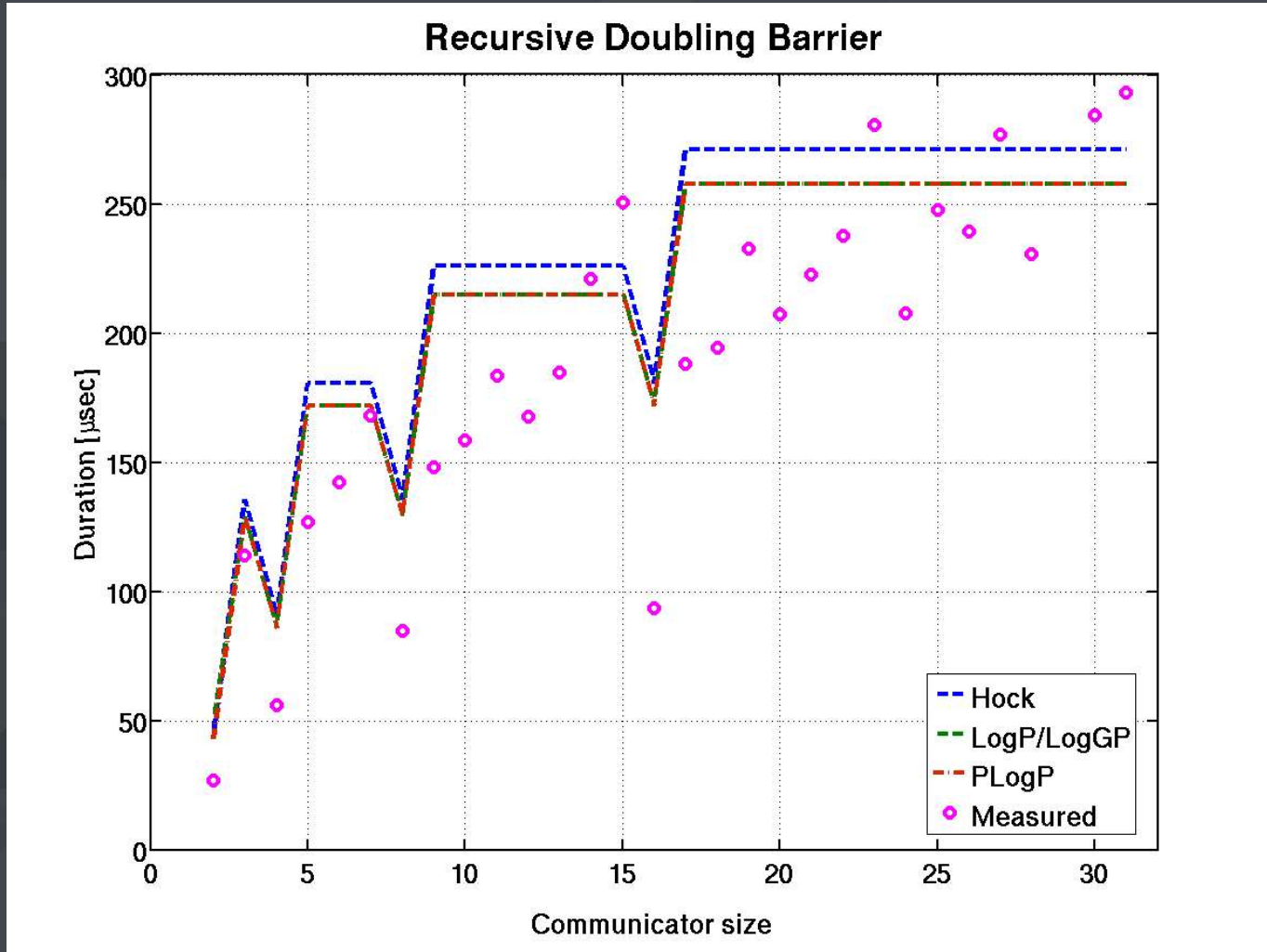
- » Built on top of Bcast and Reduce

» Assumptions

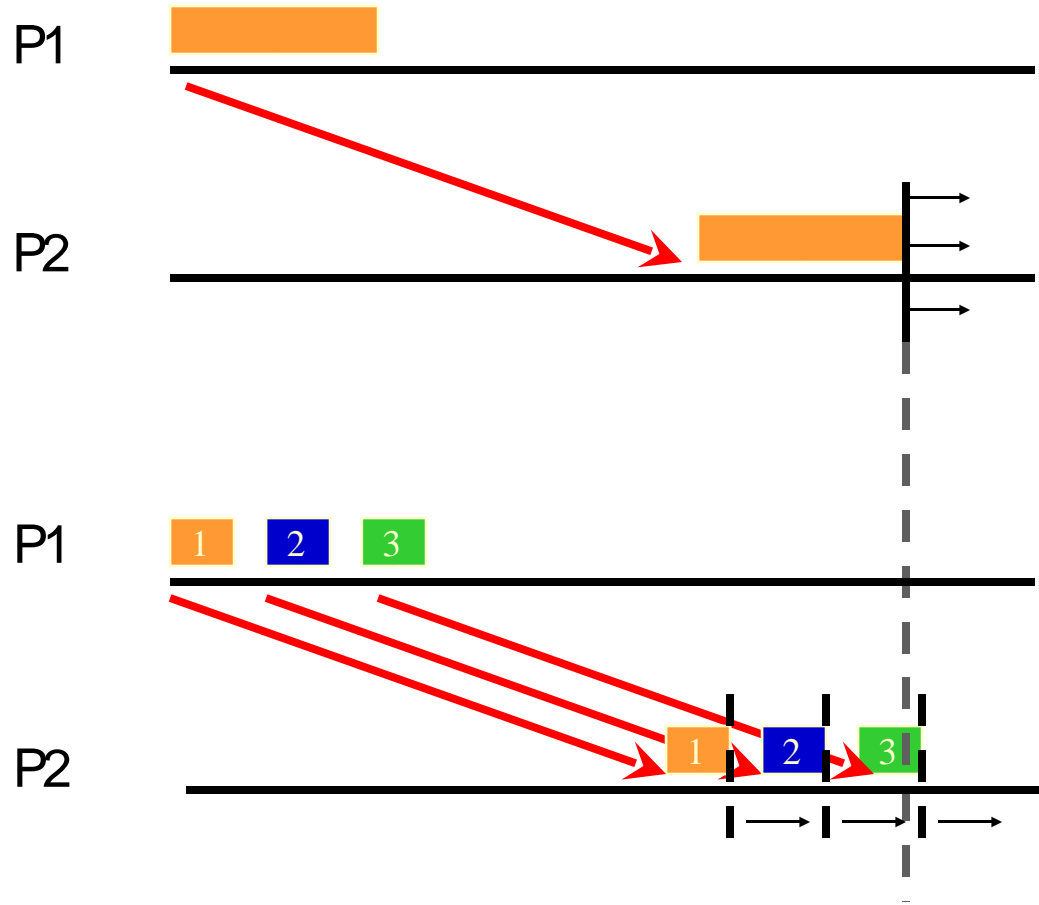
- » Fully connected and homogenous network
- » Full-duplex connection

» Binary Reduce Model

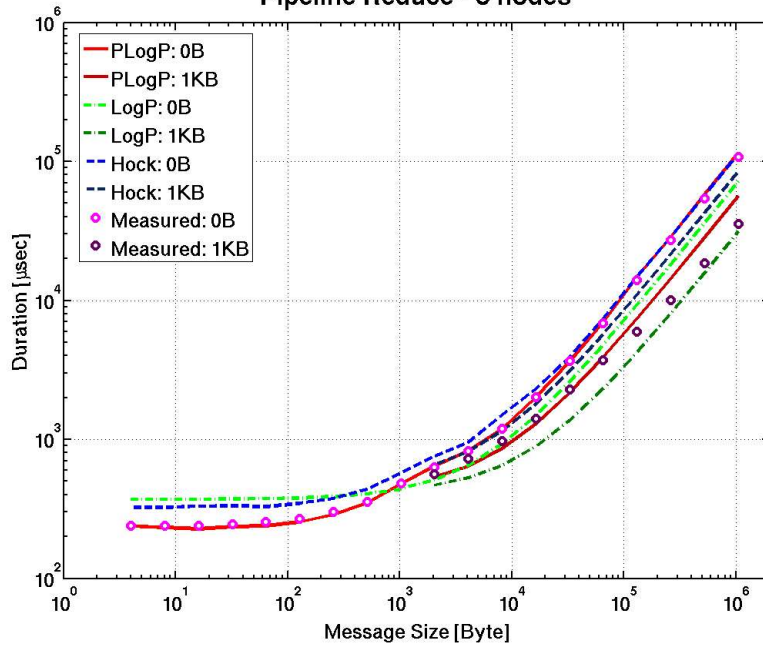
Hockney	$2 \times (\lceil \log_2(P+1) \rceil - 1) \times (\alpha(m_s) + \gamma \cdot m_s + \beta(m_s) \cdot m_s) +$ $2 \times (n_s - 1) \times (\alpha(m_s) + \gamma \cdot m_s + \beta(m_s) \cdot m_s)$
LogGP	$(\lceil \log_2(P+1) \rceil - 1) \times (L + 3 \cdot o + (m_s - 1) \cdot G + 2 \cdot \gamma \cdot m_s) +$ $(n_s - 1) \times ((m_s - 1) \cdot G + \max\{g, (3 \cdot o + 2 \cdot \gamma \cdot m_s)\})$
PLogP	$(\lceil \log_2(P+1) \rceil - 1) \times (L + 2 \times \max\{g(m_s), o_r(m_s) + \gamma \cdot m_s\}) +$ $(n_s - 1) \times (o_s(m_s) + 2 \times \max\{g(m_s), o_r(m_s) + \gamma \cdot m_s\})$



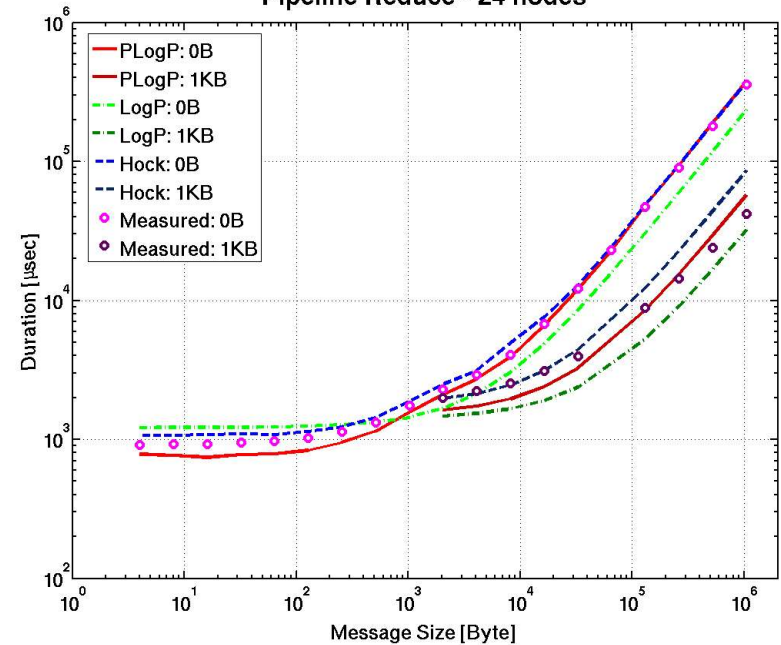
Message Segmentation

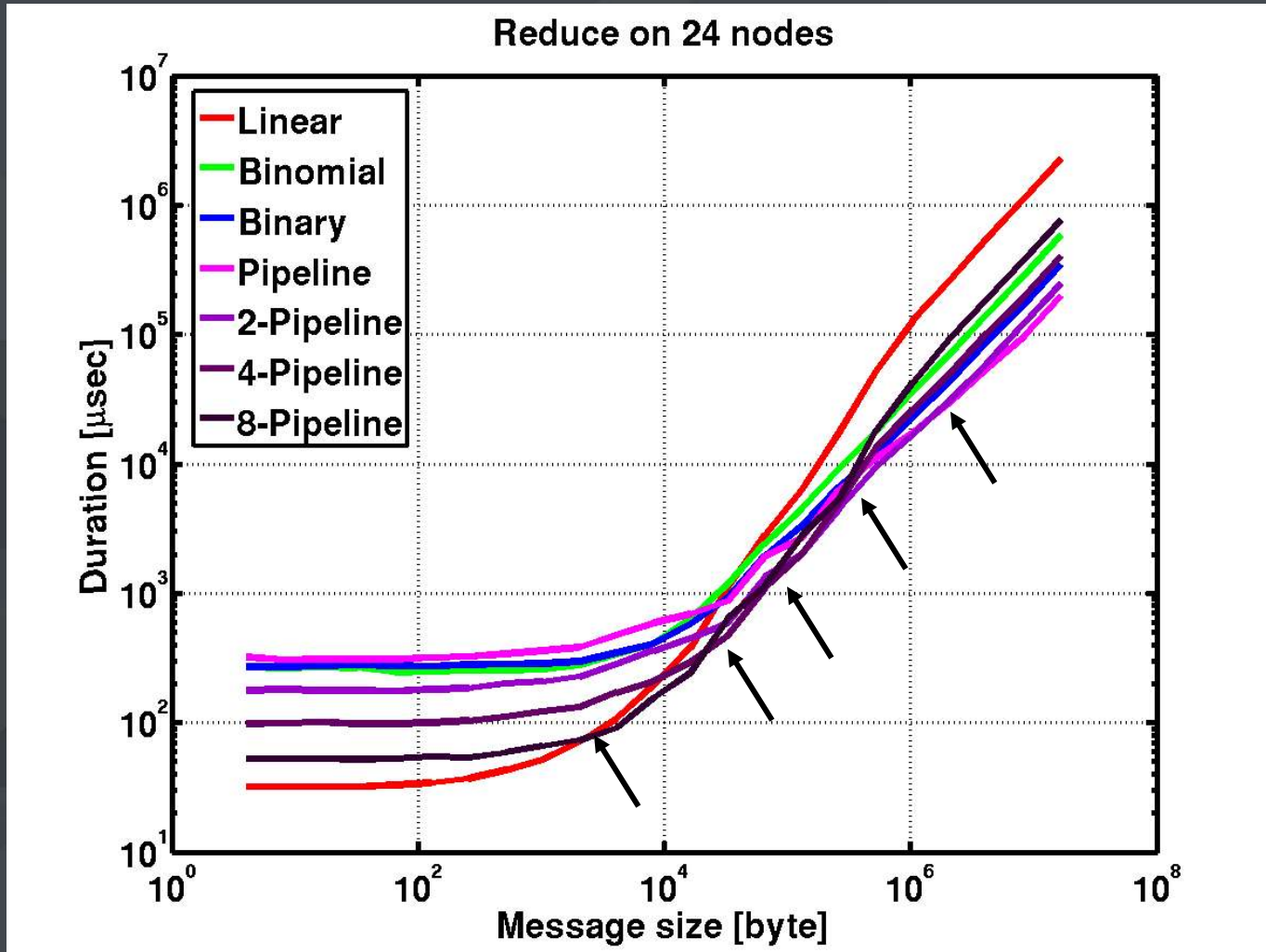


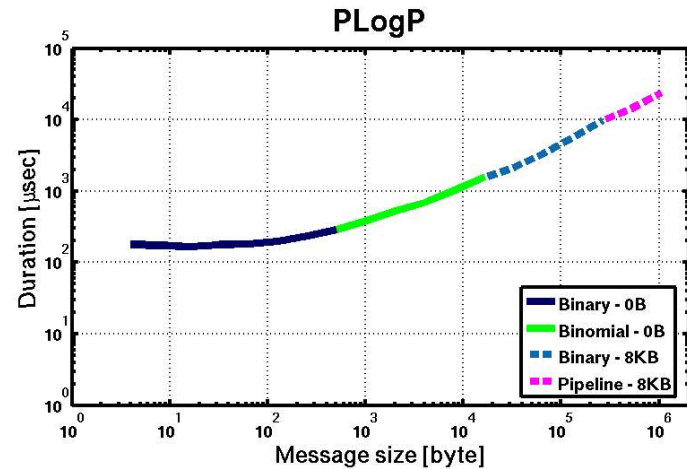
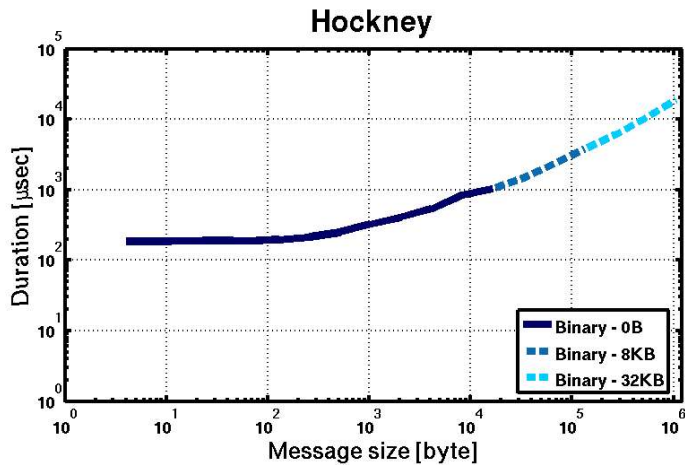
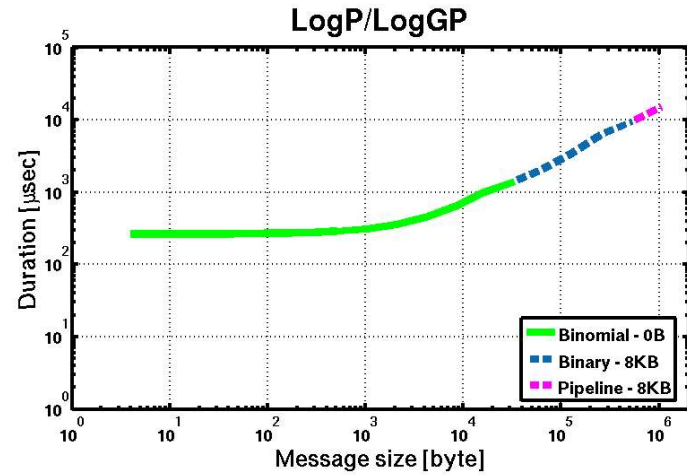
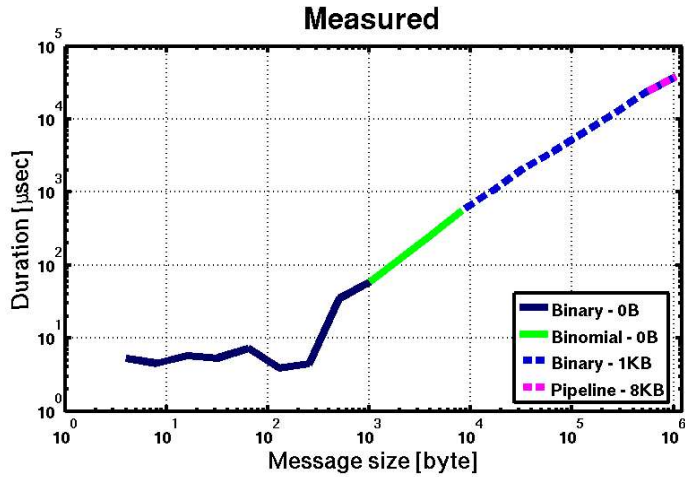
Pipeline Reduce - 8 nodes



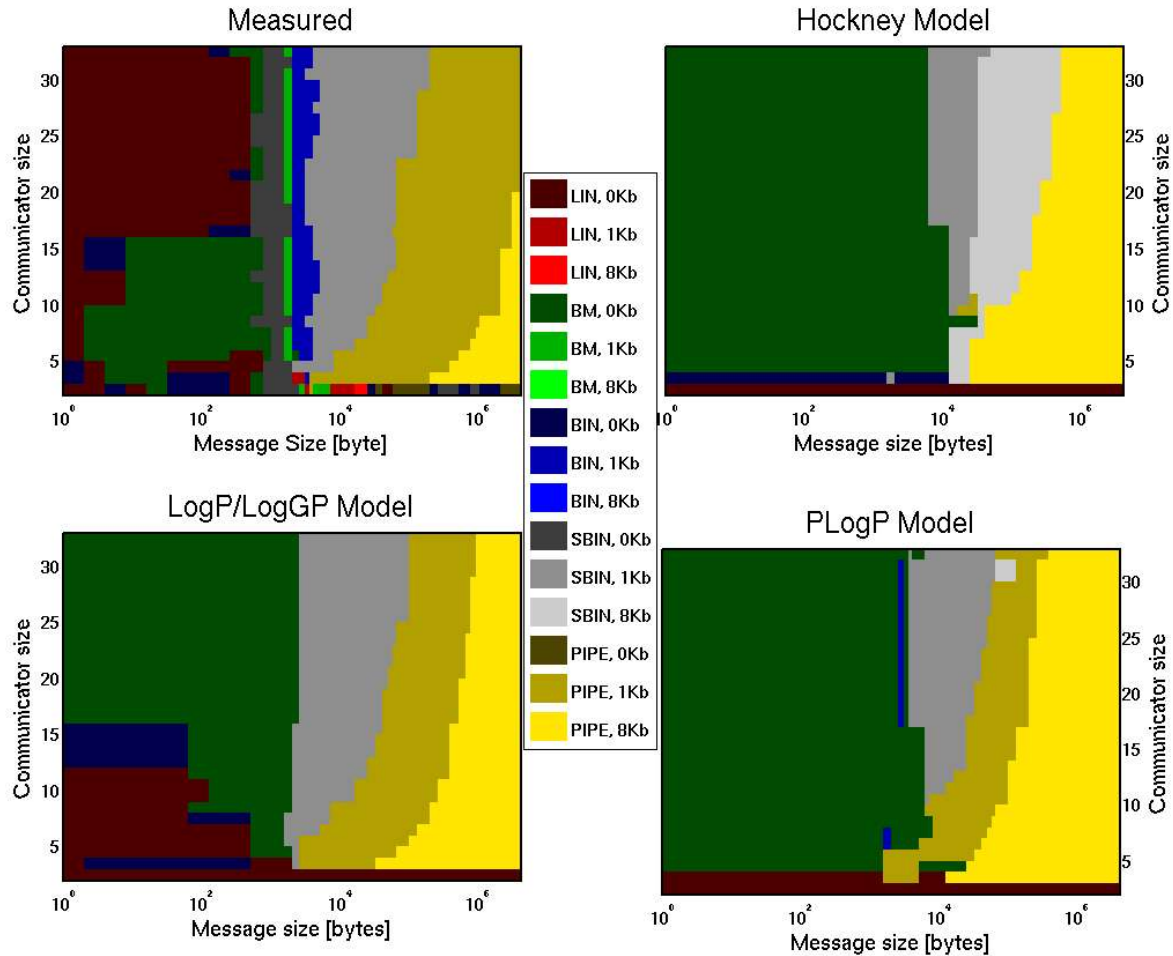
Pipeline Reduce - 24 nodes







- » Developed comprehensive set of numeric reference models for frequently used MPI collectives based on different point-to-point communication models
- » Qualitatively compared predictions of different communication models
- » Optimized FT-MPI collective operations
 - » Run time decision function
 - » Considers both communicator and message sizes
 - » Leaves room for more parameters to be considered.



- » Expanding OCC to include more collectives and more algorithms.
- » Expanding analysis to possibly include more communication and computation models.
- » Expanding work on the run-time decision functions which select optimal method for the particular collective call.
- » Incorporating this work with an automated tool to allow user to optimize collective communication on a particular system.
- » Oh, yes, what about constellations?

Thank You! :)

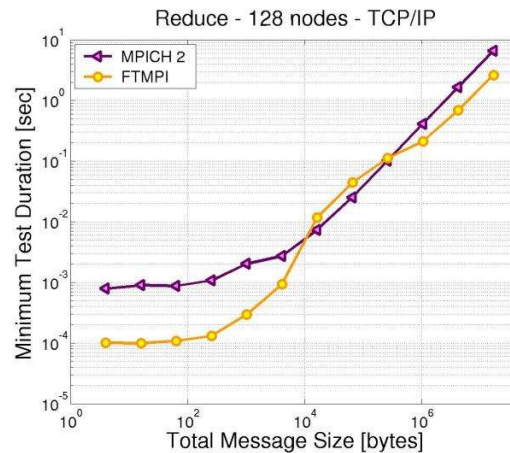
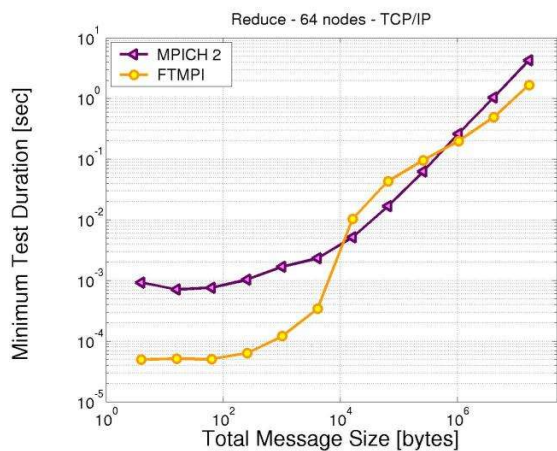
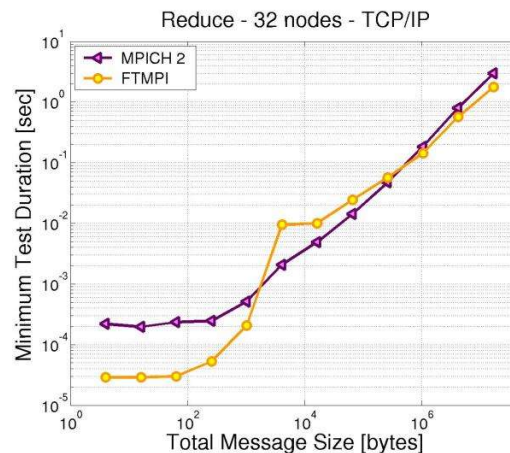
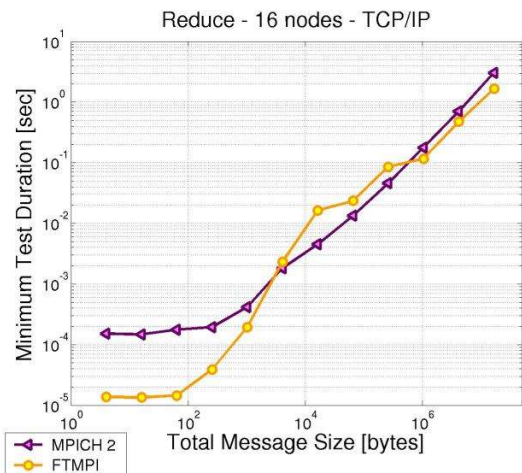
Questions?



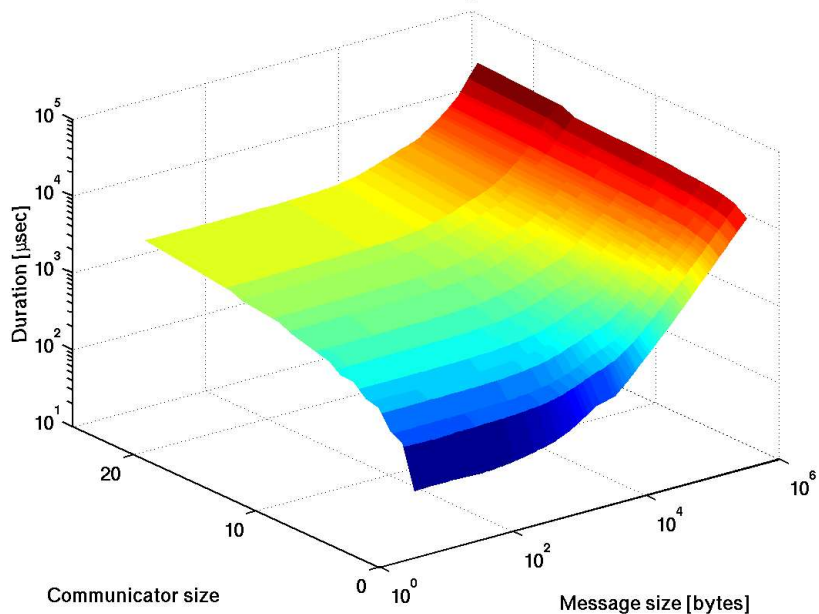
INNOVATIVE COMPUTING LABORATORY

COMPUTER SCIENCE DEPARTMENT
UNIVERSITY OF TENNESSEE

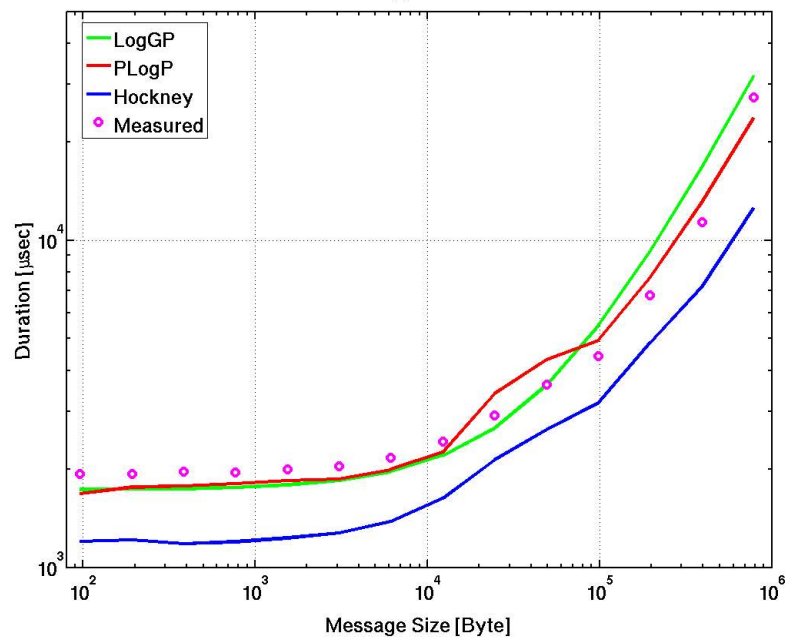
Reduce: FT-MPI vs. MPICH 2



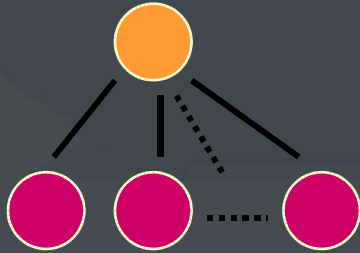
Pairwise-Exchange Alltoall



Pairwise Exchange Alltoall - 24 nodes



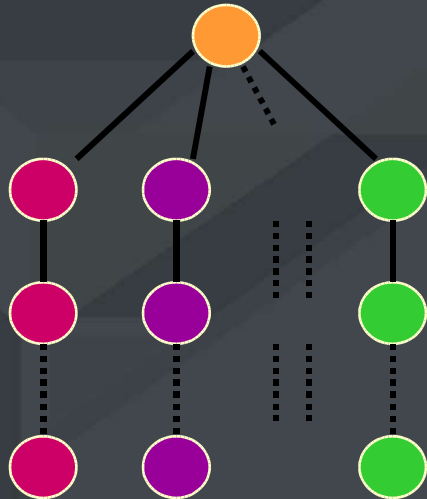
Flat tree/Linear



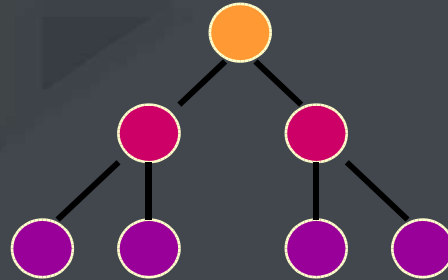
Pipeline / Ring



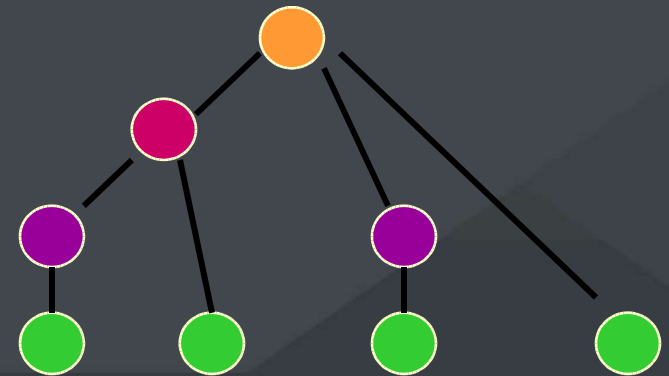
K-Chain Tree



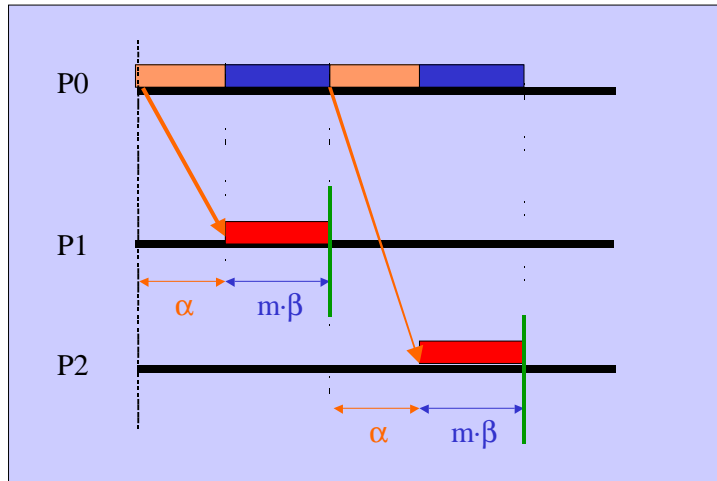
Binary tree



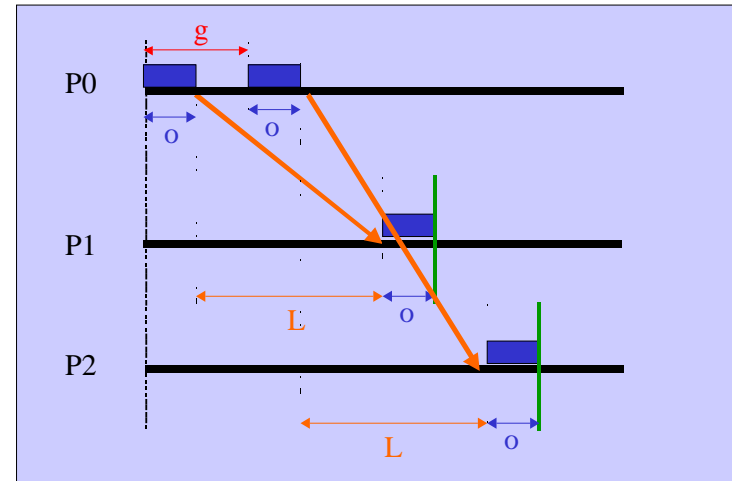
Binomial Tree



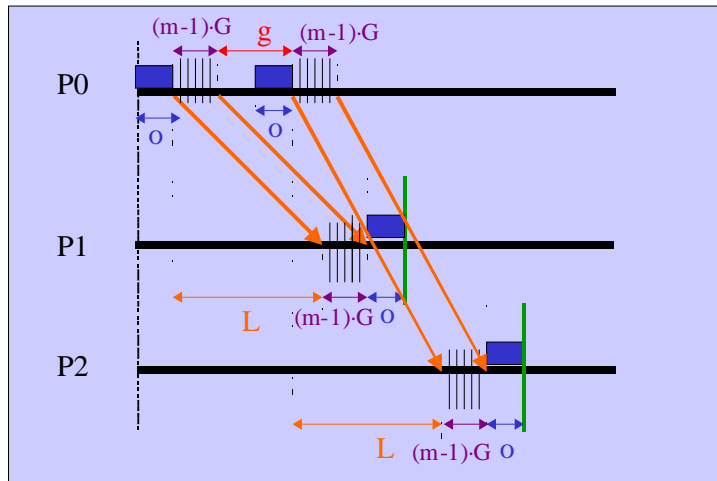
Hockney Model



LogP Model



LogGP Model



P-LogP Model

