The LogfP Model for InfiniBand™

Torsten Hoefler

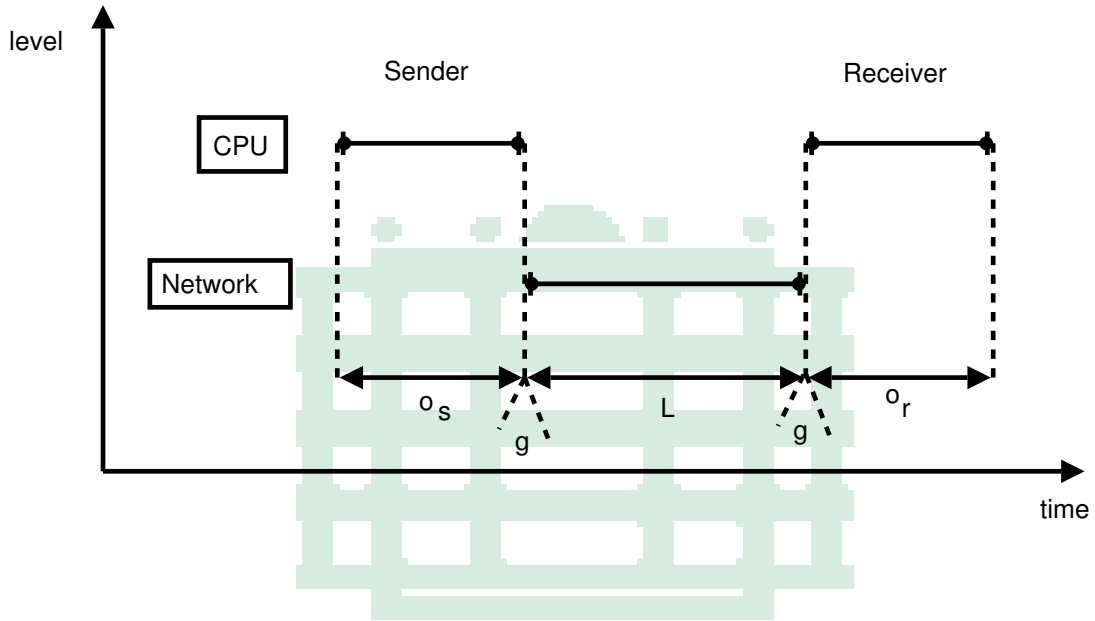
Chair of Computer Architecture
Technical University of Chemnitz

IPDPS'06 - PMEO-PDS'06 Workshop

Rhodes Island, Greece

April 29th 2006

The LogP Model



LogP Parameters

- L - Latency
- g - Gap between consecutive messages
- o - Send-/Receive Overhead
- P - Number of involved Processors

Hockney

Inaccurate but simple.

LogGP

Large messages.

pLogP

Defines parameters as function of messages size. More accurate.

LogGPS

Synchronization due to rendezvous protocol

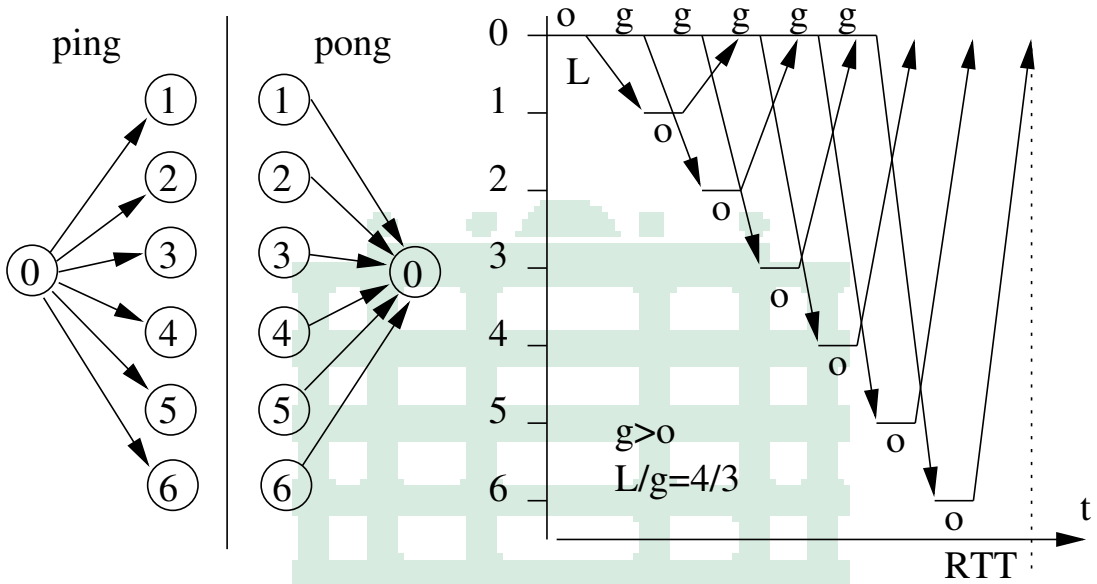
...

Many more but LogP is sufficient for small messages.

1:n n:1 Microbenchmark

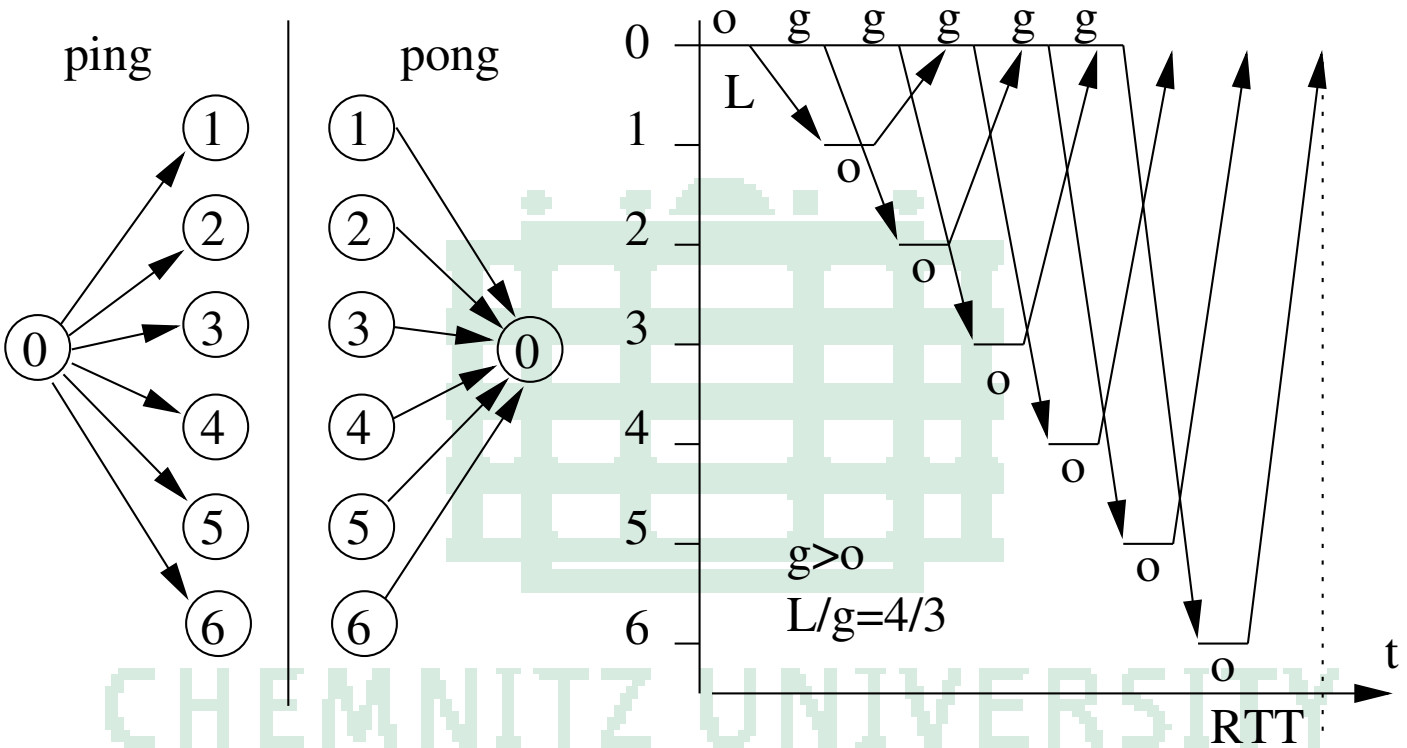
- Developed to analyze the InfiniBand™ network
- Especially for collective communication
- One to many communication
- Measures single message performance (RDTSC)
- MPI based
- Supports (nearly) all transport types

1:n n:1 Microbenchmark - principle



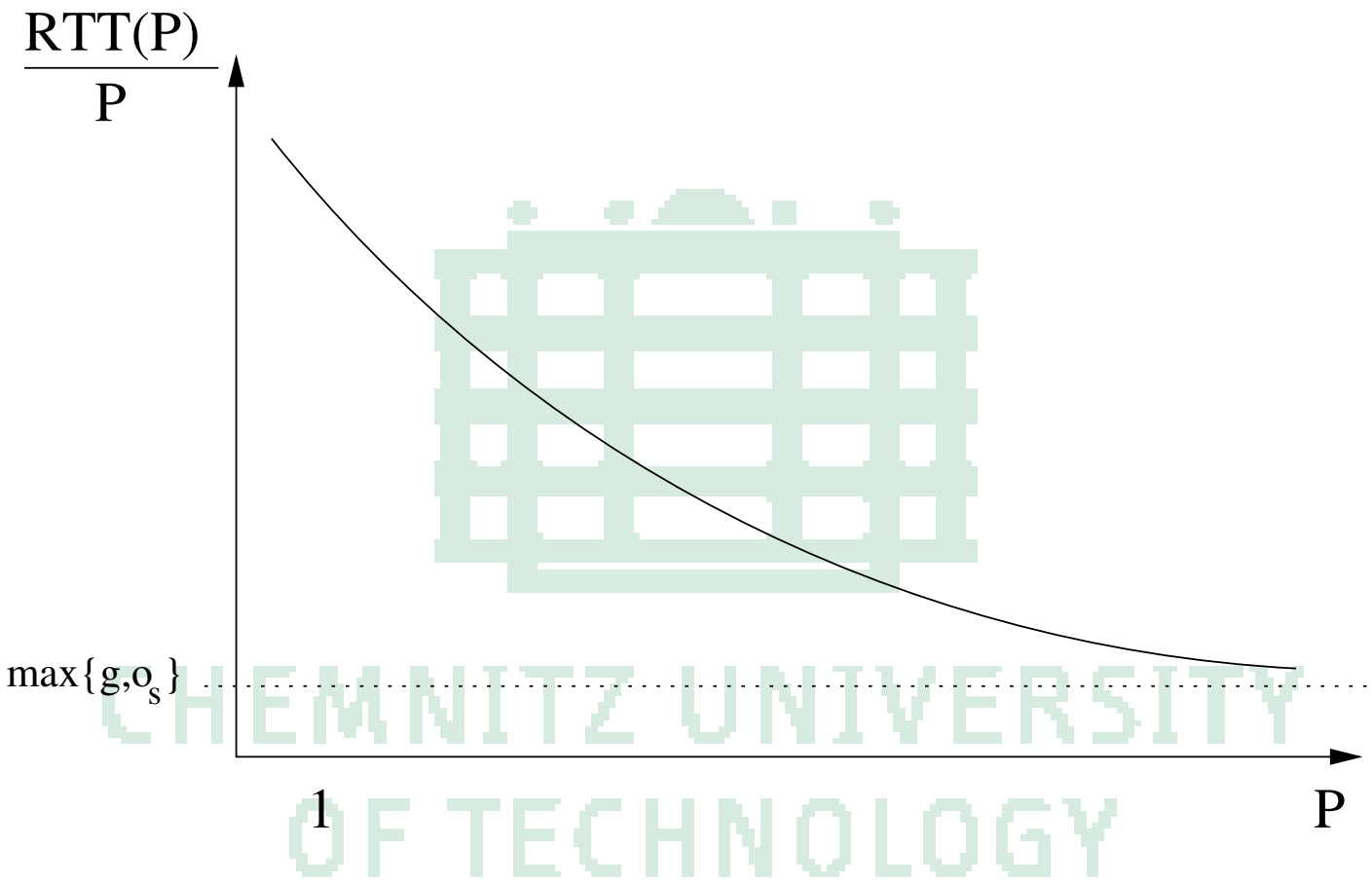
1. (0): Take time
2. (1..n-1): Send a single message to n-1 hosts
3. (1..n-1): Hosts respond immediately
4. (0): Wait for message receiption from all hosts
5. (0): Take time

LogP Performance Prediction



1. special case RDMA/W ($o_r = 0$)
2. $RTT(P)/P = (2o + 2L + (P-1) \cdot max\{g, o\})/P$

LogP Performance Prediction - RTT



- $RTT(P)/P = (2o + 2L + (P-1) \cdot max\{g, o\})/P$

Test Systems

Opteron PCI-X

- Opteron with PCI-X InfiniBand™
- Thanks to T. Klug and C. Trinitis from Technical University of Munich

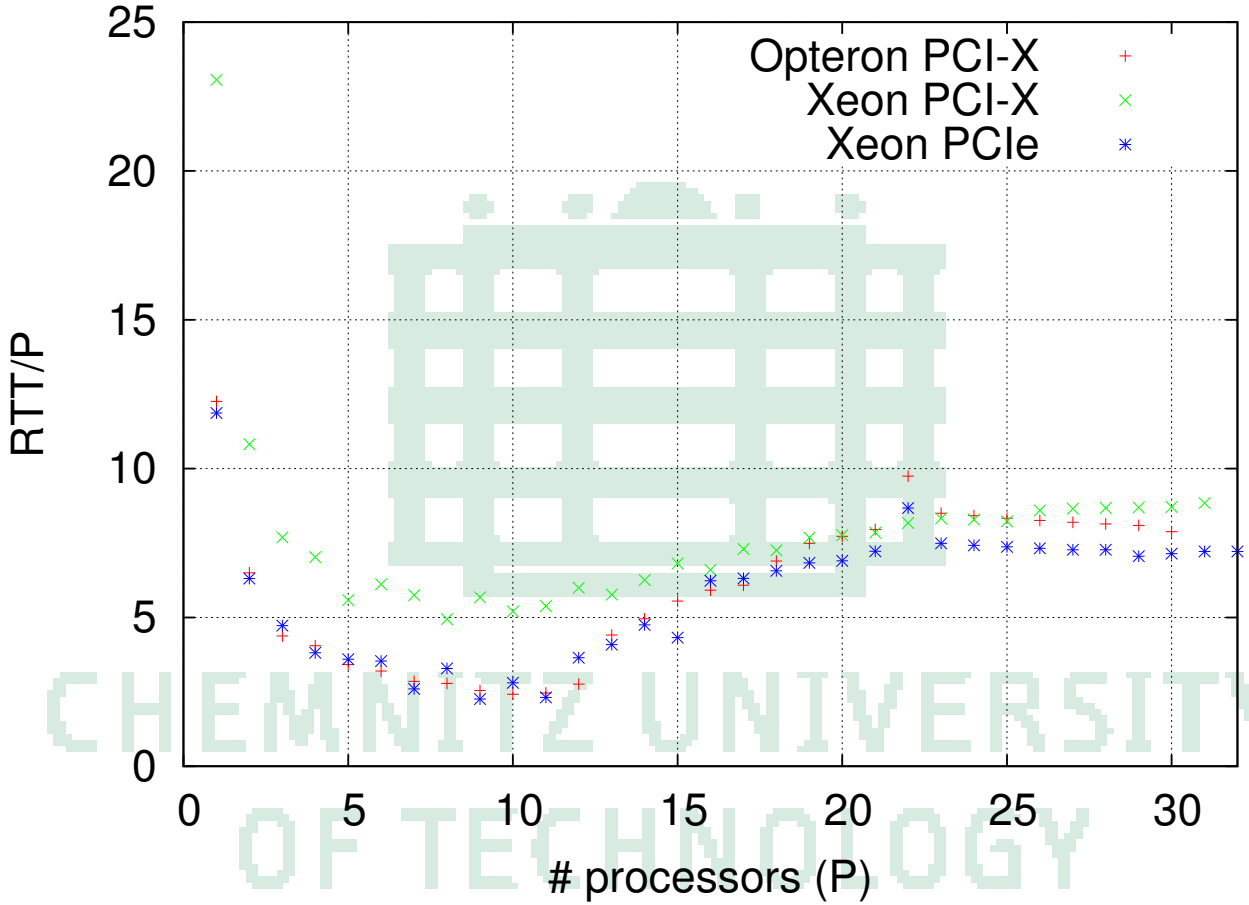
Xeon PCI-X

- Xeon with PCI-X InfiniBand™
- Thanks to University of Stuttgart

Xeon PCI-e

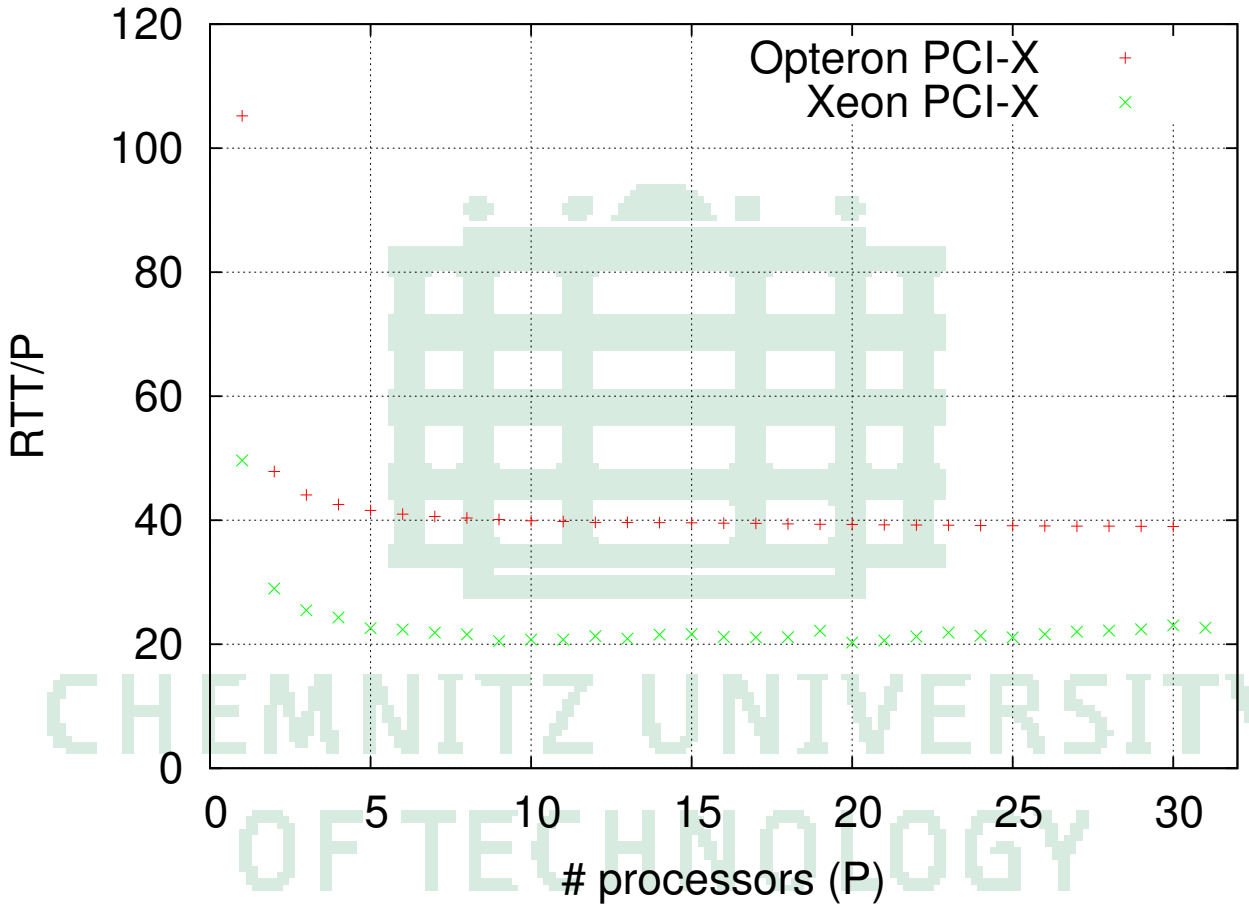
- Xeon with PCI-e InfiniBand™
- Thanks to J. Simon from University of Paderborn

1:n n:1 Benchmark Results



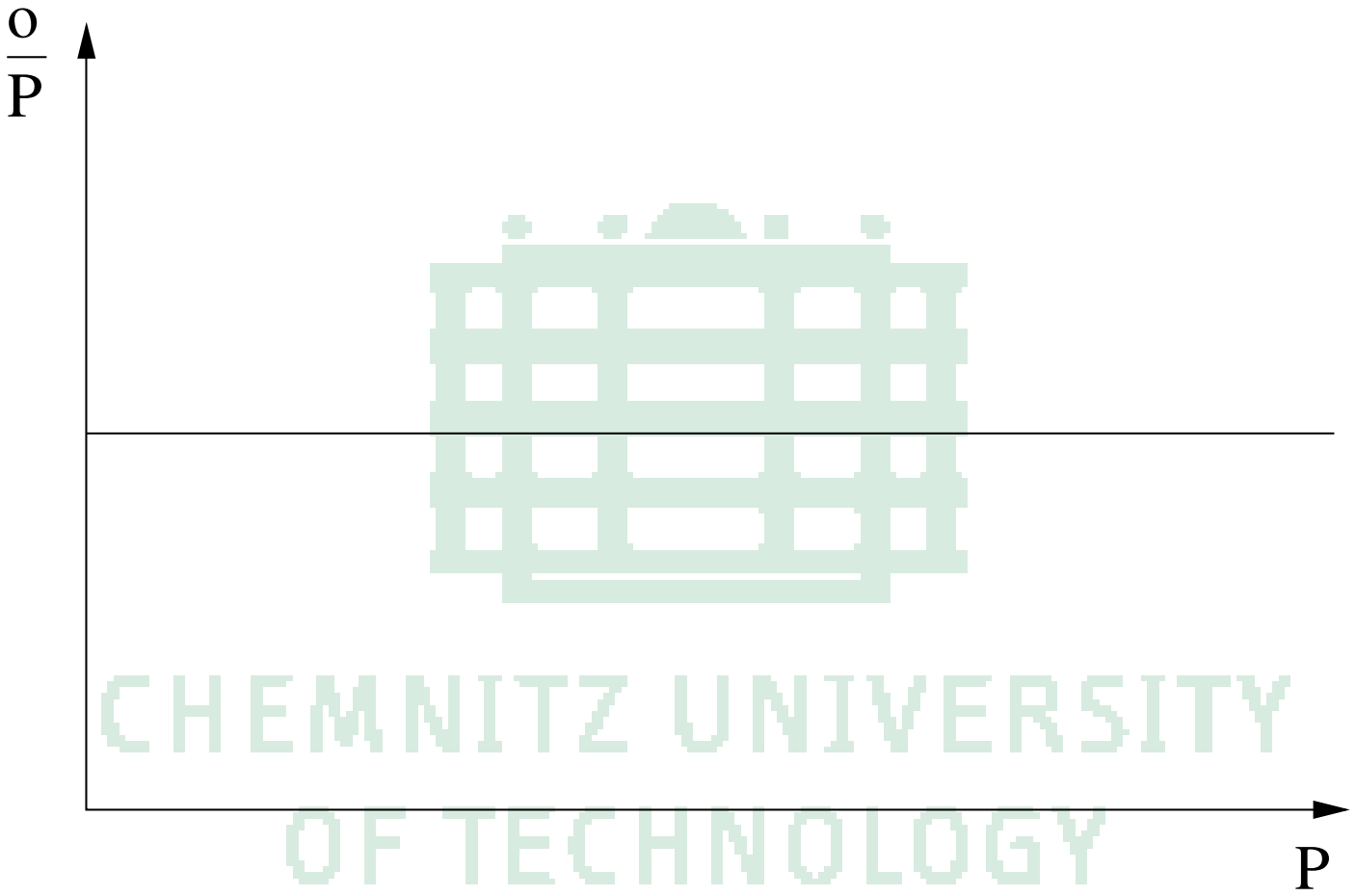
- $RTT(P)$ for 1 byte

1:n n:1 Benchmark Results



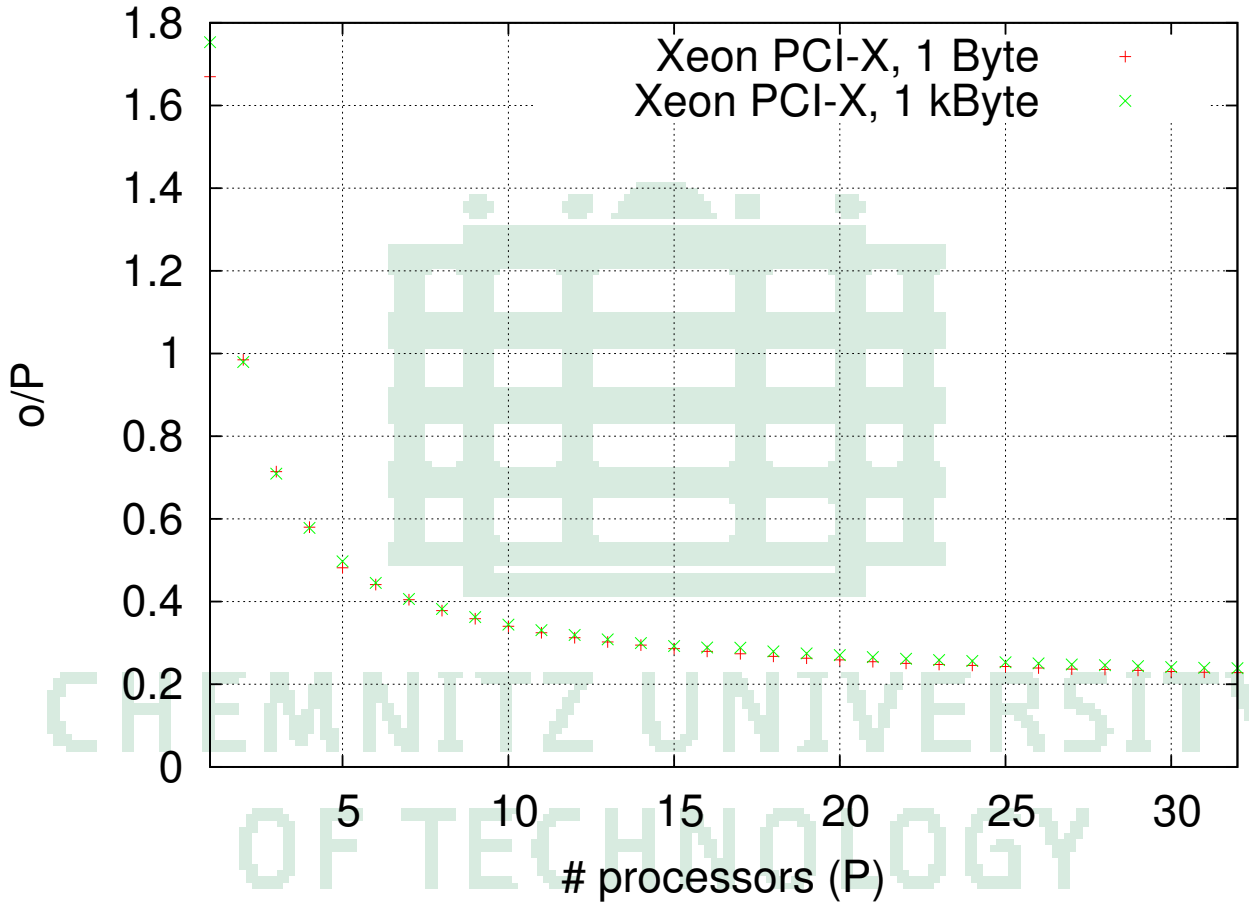
- $RTT(P)$ for 8 kbyte

LogP Performance Prediction - o



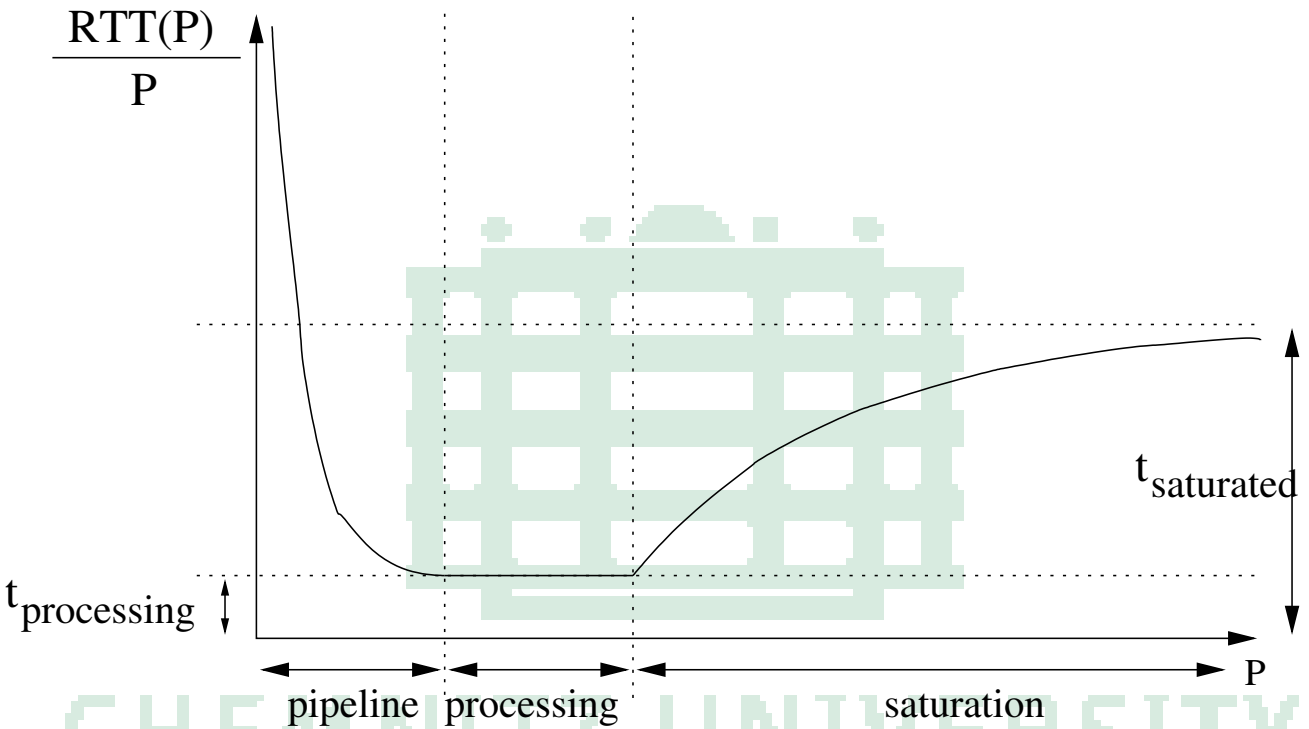
- *o* is constant in LogP

1:n n:1 Benchmark Results



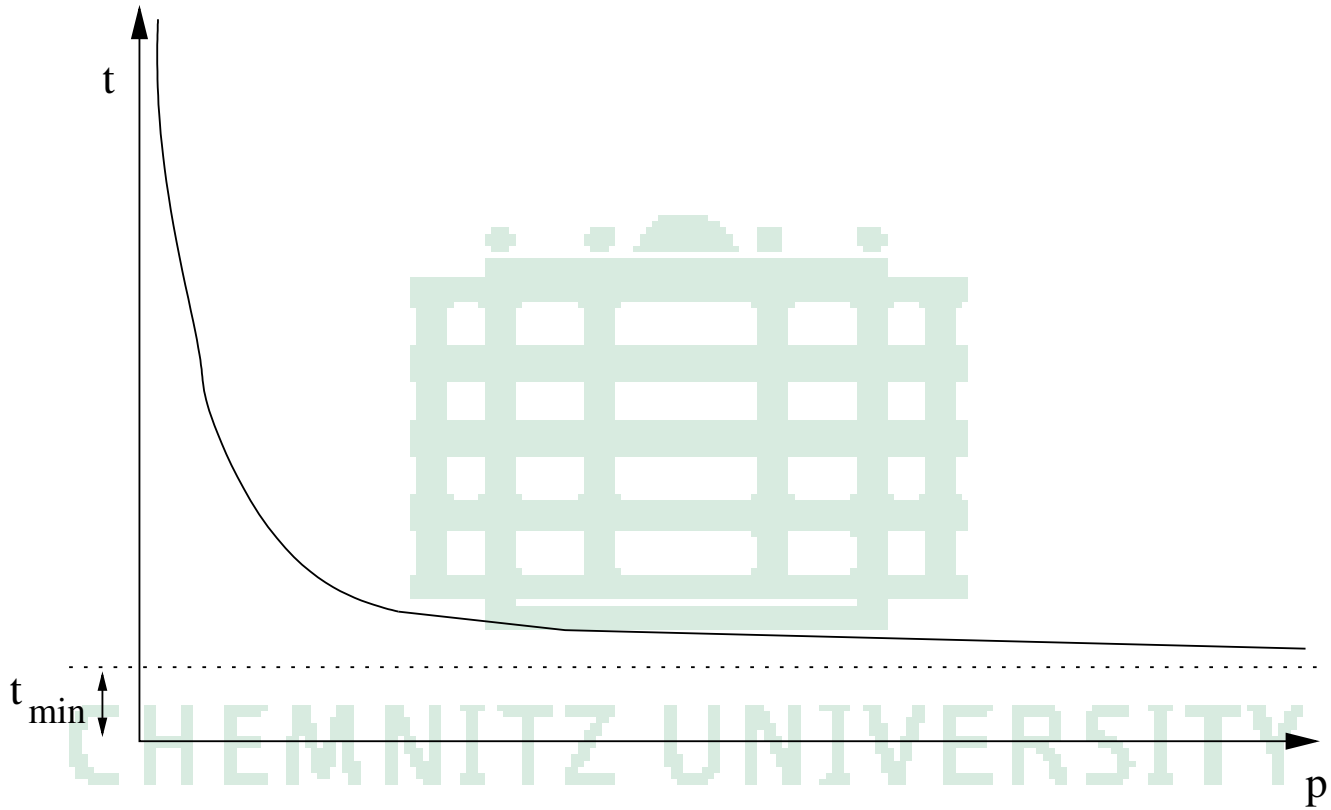
- *o* for 1 byte

RTT Model



- $RTT = t_{pipeline} + t_{processing} + t_{saturation}$
- 6 nonlinear parameters (*e*-function)
- very accurate but unuseable

o Model



- $RTT = t_{pipeline}$
- 3 nonlinear parameters ($1/x$-function)
- very accurate but unuseable

The LogfP Model

Goals

- better than plain LogP
- accurate and simple
- less parameters

Results

- based on empiric analysis
- *f*-parameter denotes ”free” messages
- surprisingly accurate

The LogfP Model

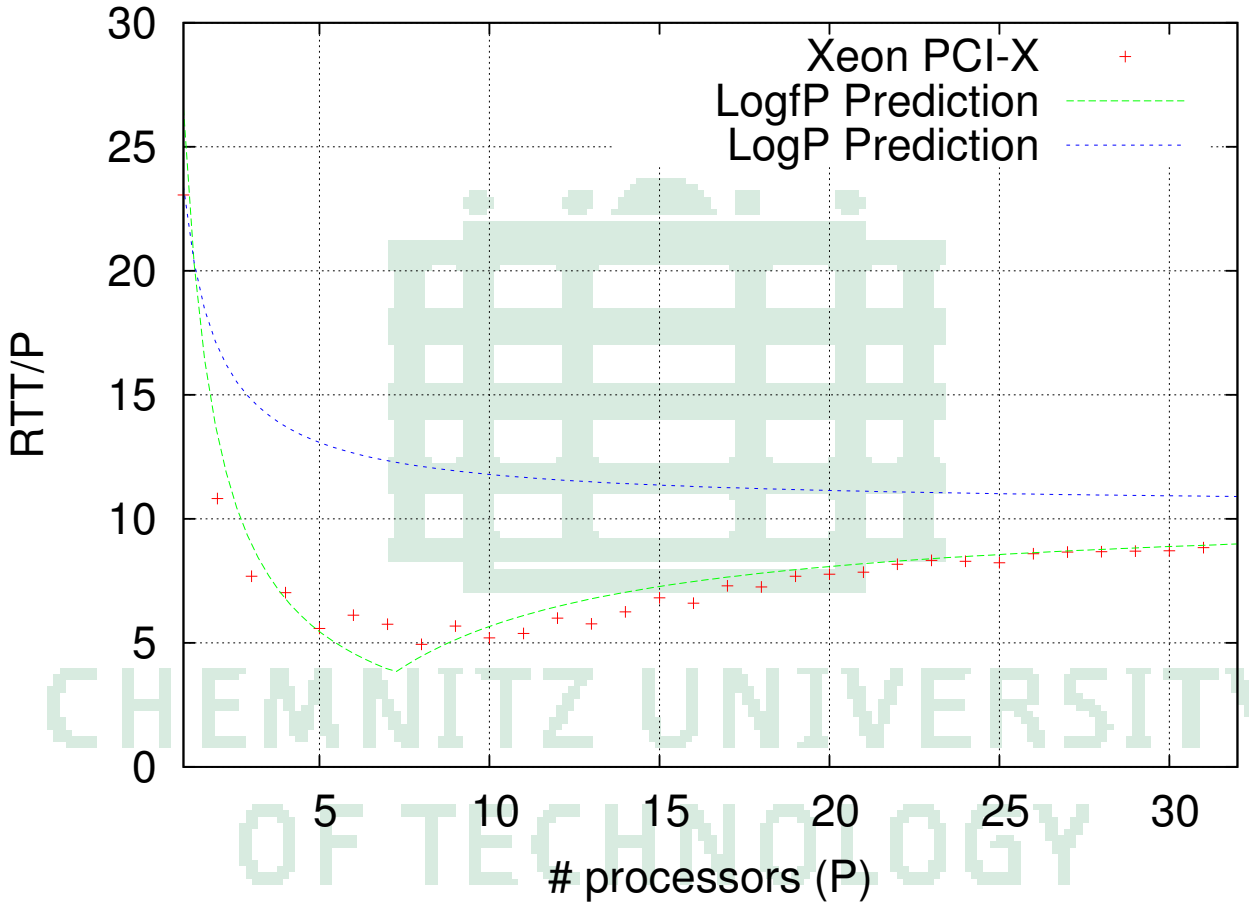
Overhead

- $o(P) = o_{min} + o_{max}/P$
- $o_{min} = o(P \to \infty)$
- $o_{max} = o(P \to 1)$

RTT

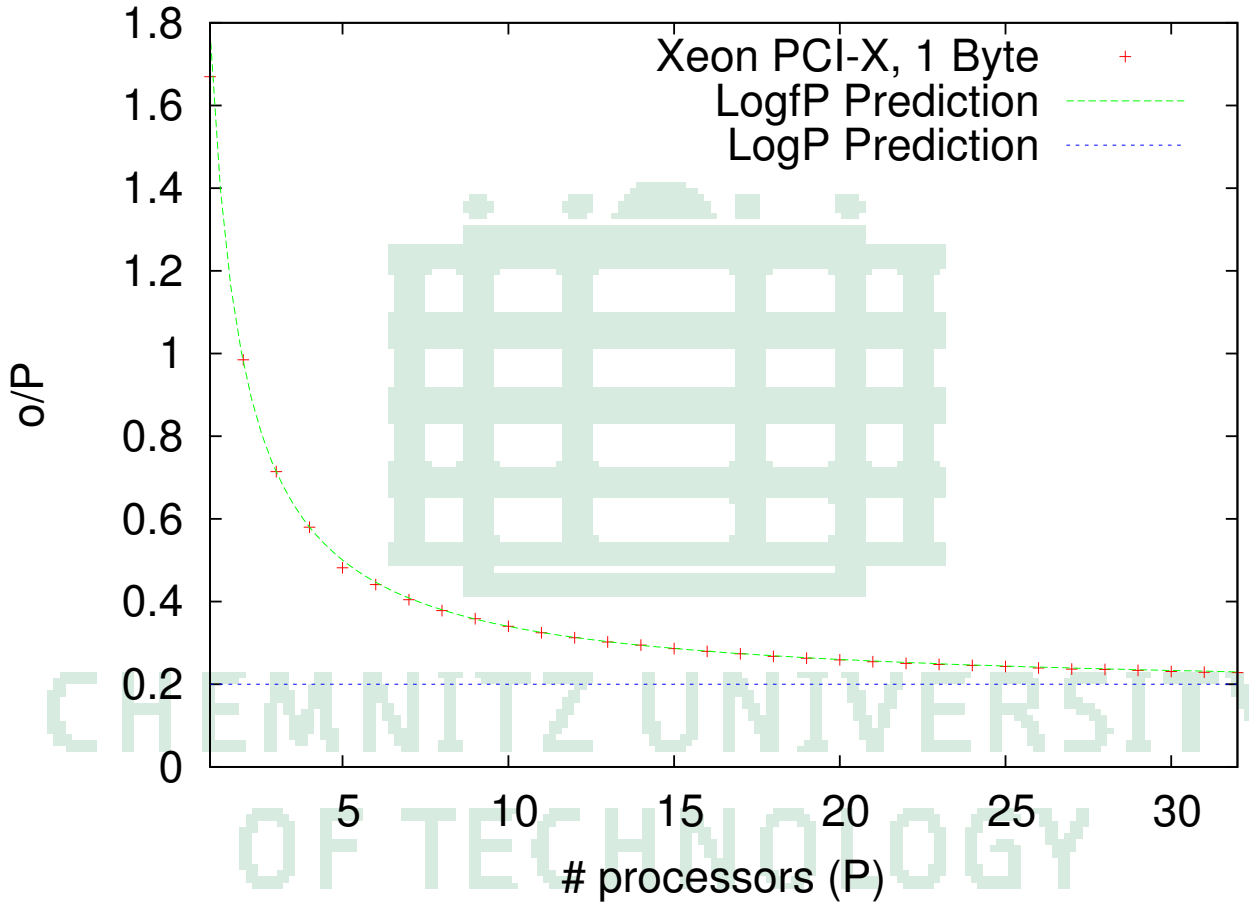
- f denotes number of ”free” messages (minimum in RTT(P)/P Graph)
- $\forall(P \leq f)\ RTT(P) = 2L + P \cdot o_s(P) + o_s(1)$
- $\forall(P > f)\ RTT(P) = 2L + o(P) + o_s(1) + max\{(P-1) \cdot o(P), (P-f) \cdot g\}$

RTT Model



- $f = 7$

o Model



- $o_{min} = 0.18, o_{max} = 1.6$

Summary

Comparison to LogP

- simple extension
- more accurate than LogP
- less accurate as LoP but useable

Achievements

- optimized InfiniBand™ barrier (see CAC'06 Workshop)

Future Work

- more InfiniBand™ optimized collectives for Open MPI
- detailed analysis for large messages