Effects of Communication Latency, Overhead, and Bandwidth in a Cluster Architecture

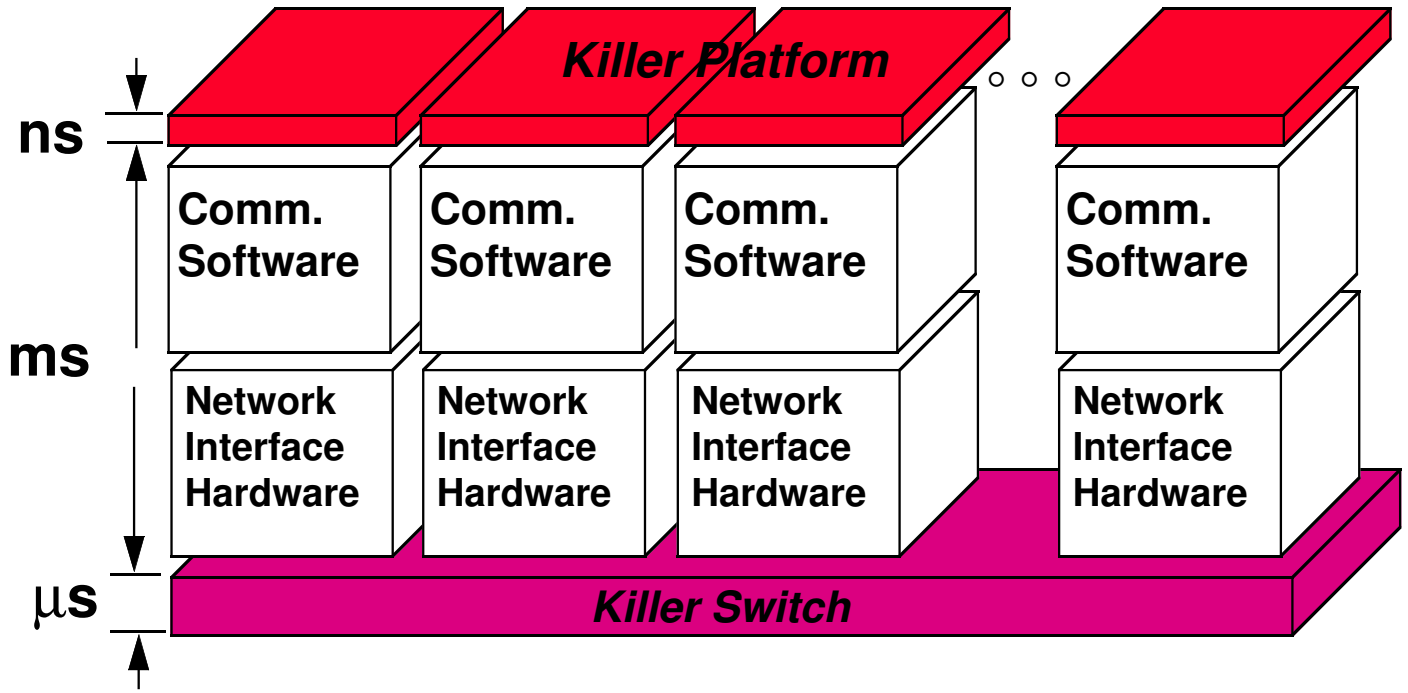
David E. Culler

Rich Martin, Amin Vahdat

Computer Science Division

University of California, Berkeley

The Fast Communication Problem

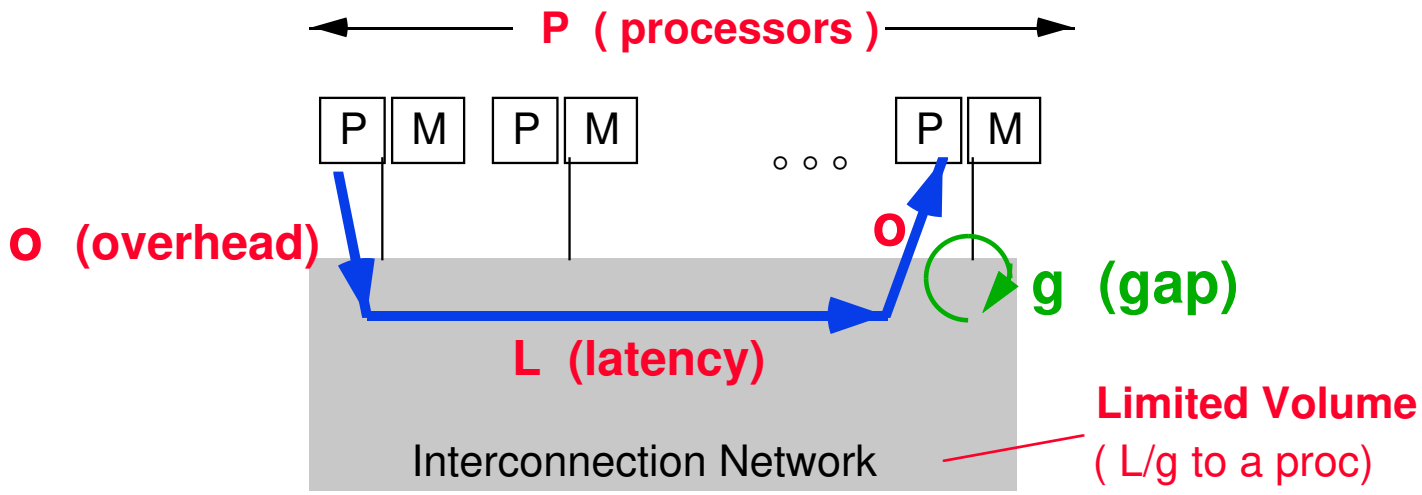


- **NOW Goal: 10 µs end-to-end in general purpose, high-performance clsuters**

Question of the Study

- How much do our efforts really improve application performance?
- Have we gone far enough?
- To what aspects of communication performance are applications most sensitive?
- Can we predict it?

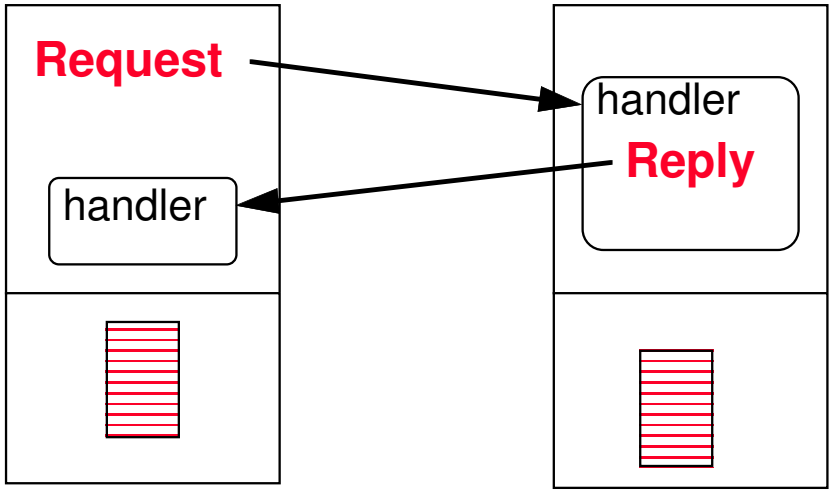
Model Framework: LogP



- **Latency in sending a (small) message between modules**
- **overhead felt by the processor on sending or receiving msg**
- **gap between successive sends or receives (1/rate)**
- **Processors**

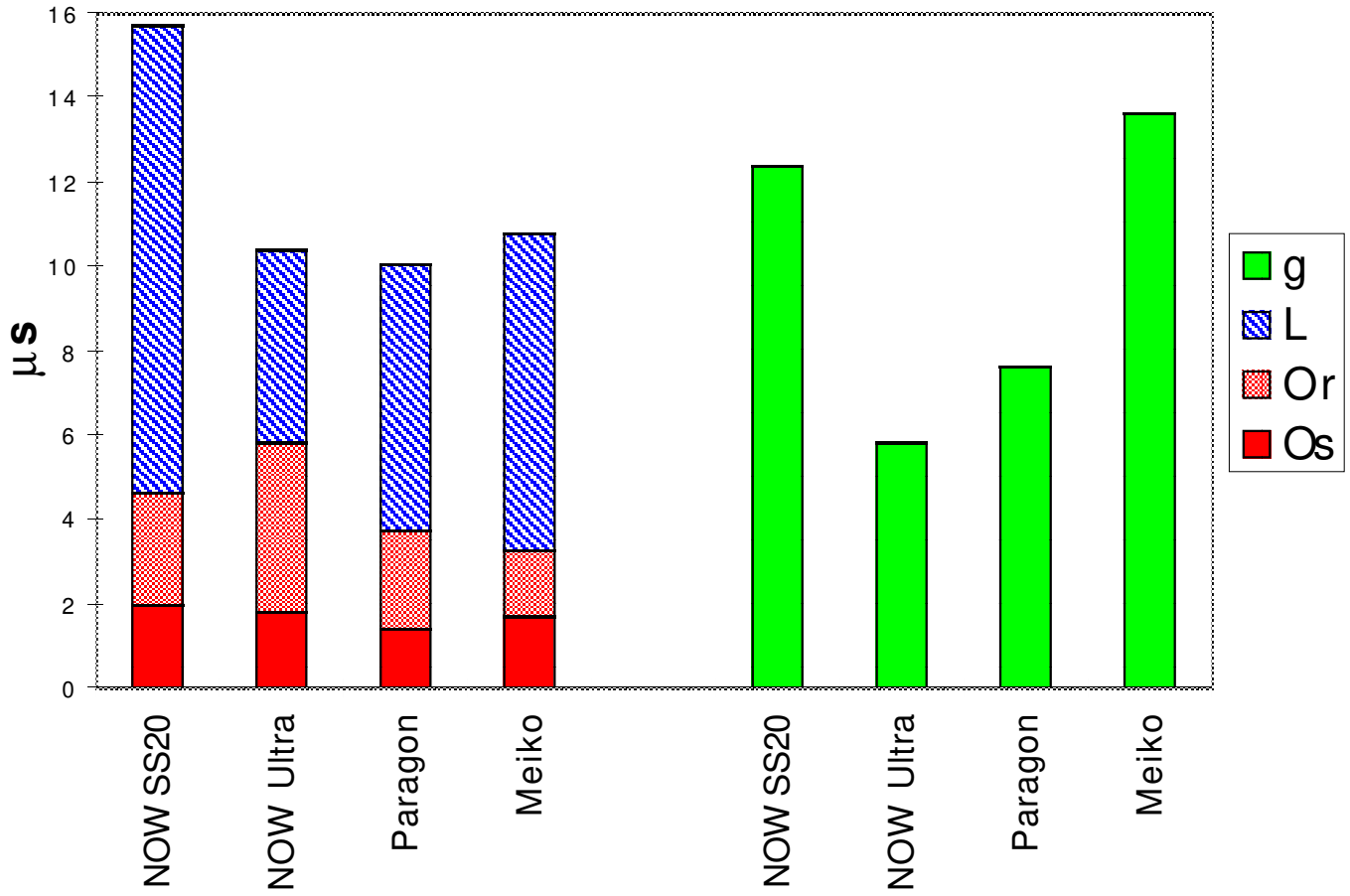
Generic Active Messages (GAM)

- **Highly optimized communication layer on a range of HW**
- **Request / Reply small active messages (RPC)**
- **Bulk-Transfer (store & get)**



Round Trip time: 2 x (2o + L)

LogP Summary of Current Machines



Max MB/s: 38 141 47

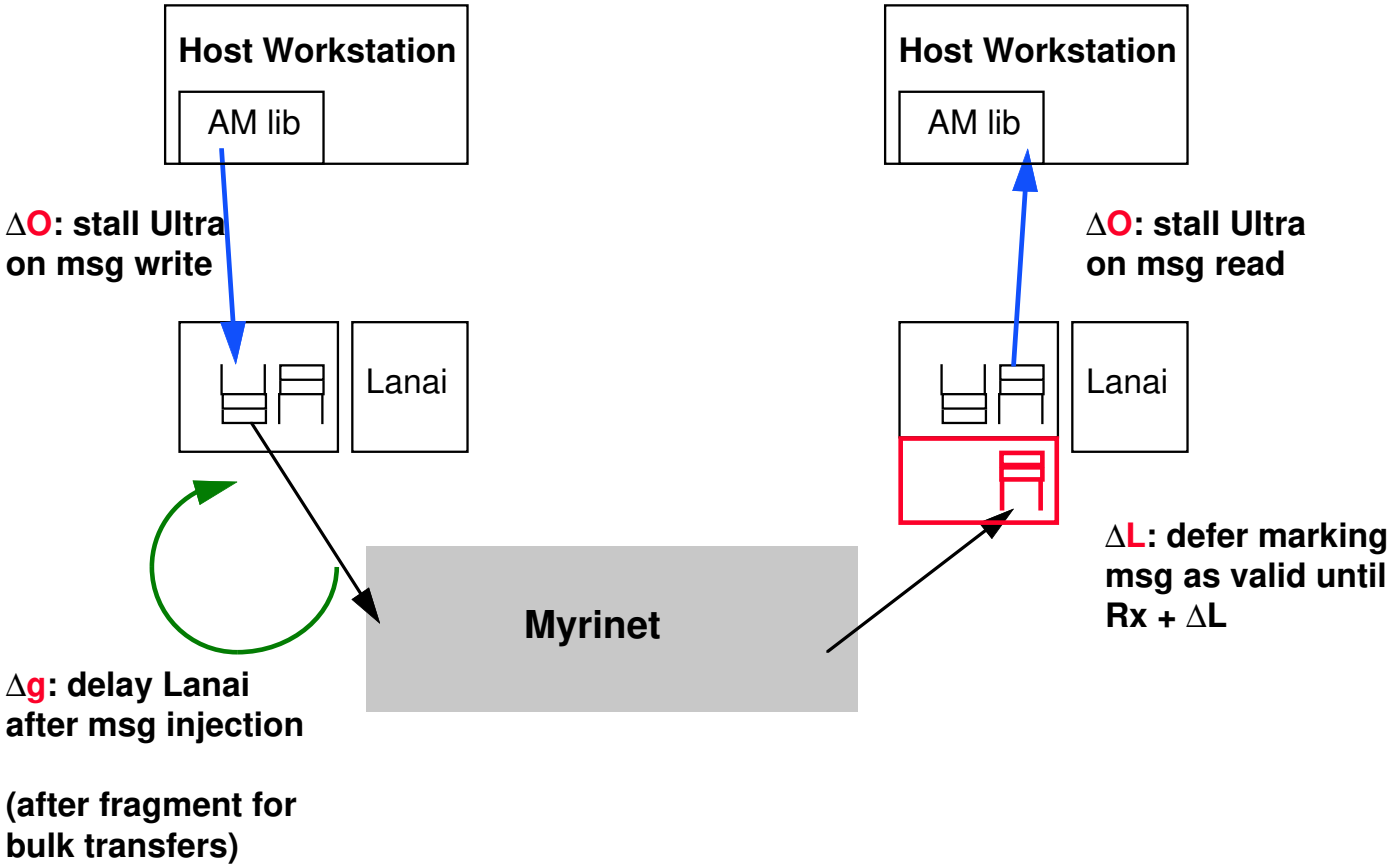
Outline of the Talk

- Framework for Communication Performance
- Empirical Method
- Base Application Characteristics
- Sensitivity to Overhead
- to gap
- to Latency
- to bulk transfer bandwidth
- Predictive models
- Conclusions

Methodology

- **Apparatus:**
  - **35 Ultra 170s (64 MB, .5 MB L2, Solaris 2.5)**
  - **M2F Lanai + Myricom Net in Fat Tree variant**
  - **GAM + Split-C**
- **Modify the Active Message layer to inflate L, o, g, or G independently**
- **Execute a diverse suite of applications and observe effect**
- **Evaluate against natural performance models**

Adjusting L, o, and g (and G) *in situ*



Microbenchmarks for Calibrating the Empirical Adjustments

For Range of Δ and *M*

Repeat *M* times

Issue request

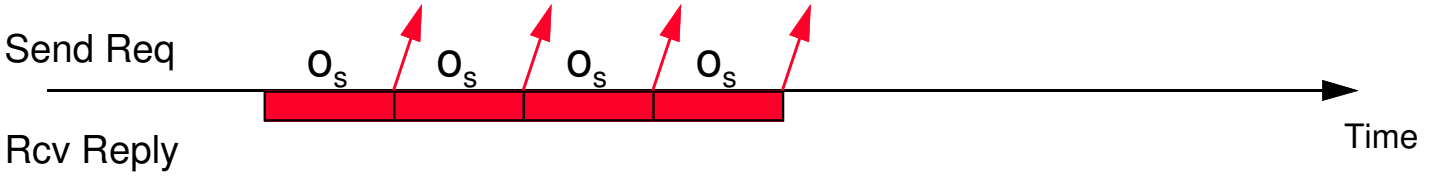
Compute for Δ

Time(*M*, Δ)

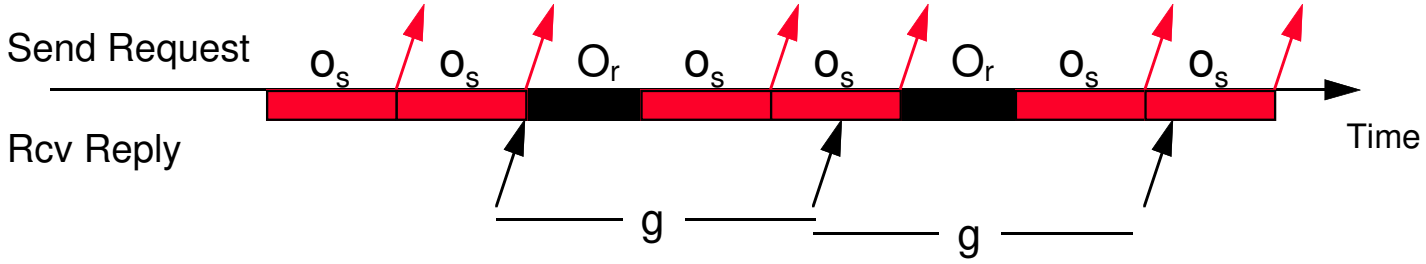
... things complete

Time to Issue Sequence

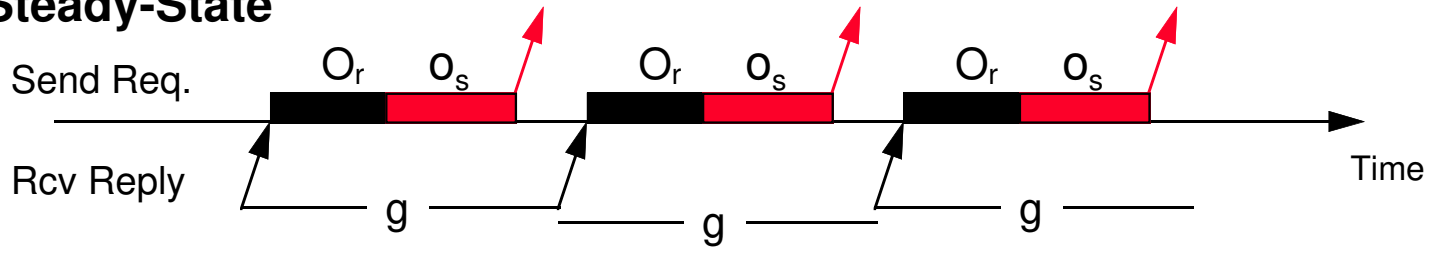
Send Only



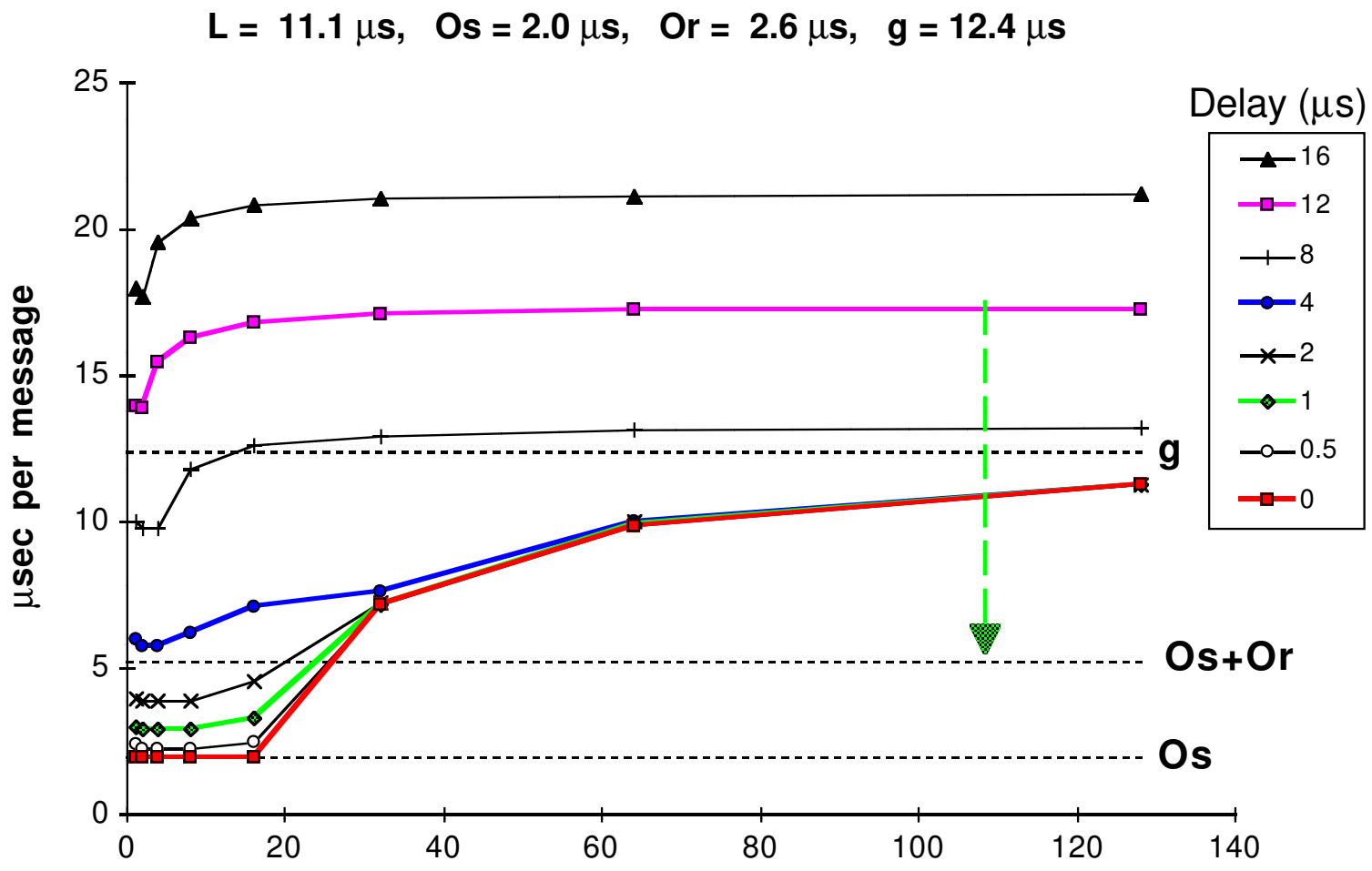
Transition



Steady-State



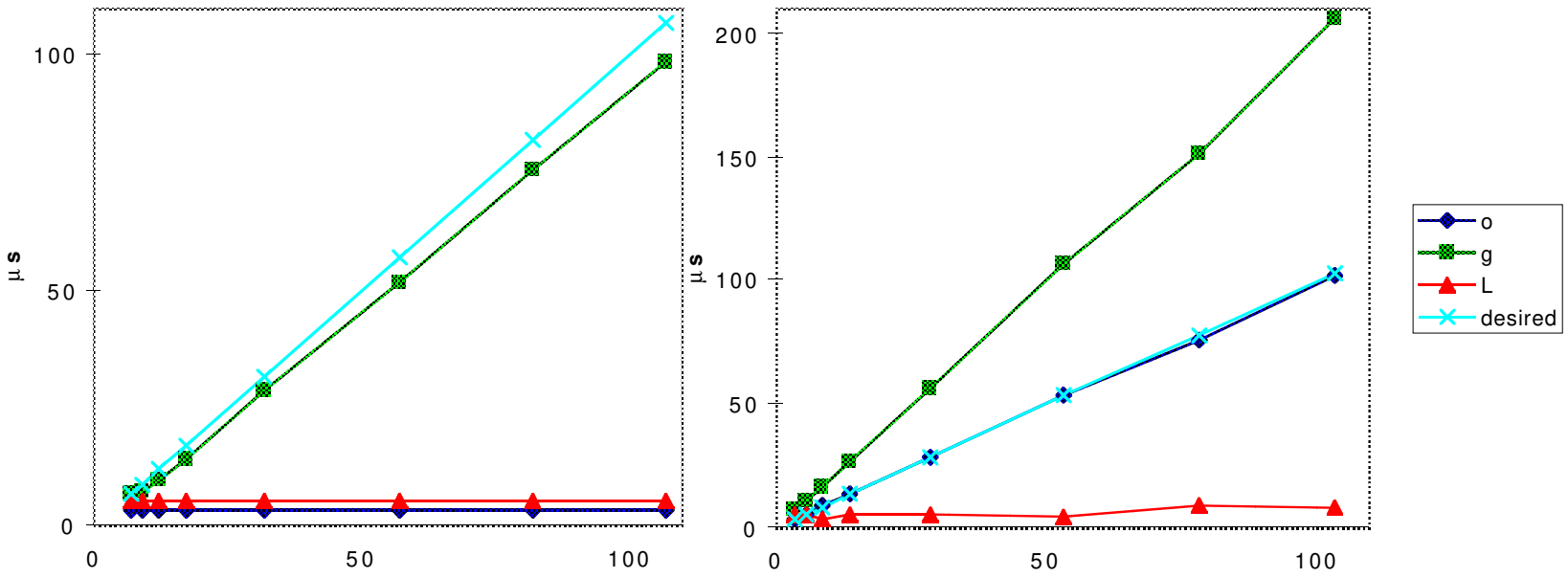
NOW-SS20 Microbenchmark Signature



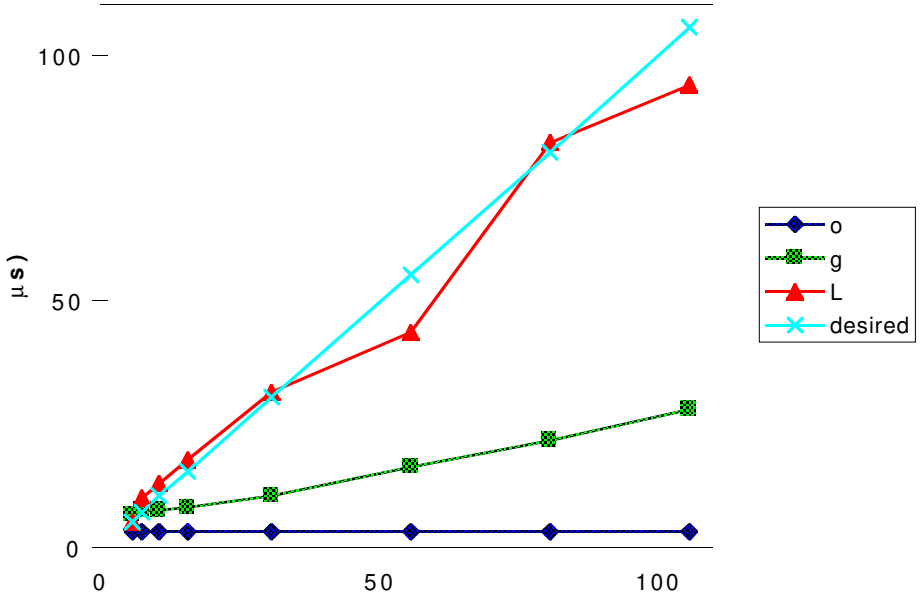
Credit: Richard Martin

LogP Sensitivity 12

Calibration



Calibration (cont)



Applications Characteristics

- **Message Frequency**
- **Write-based vs. Read-based**
- **Short vs. Bulk Messages**
- **Synchronization**
- **Communication Balance**

Applications used in the Study

Program		Input	P=16	P=32	Msg Interval (µs)
Radix	Integer radix sort	16M 32-bit keys	13.7	7.8	6.1
EM3D(write	Electro-magnetic wave	80K Nodes, 40% rmt	88.6	38.0	8.0
	Electro-magnetic wave	80K Nodes, 40% rmt	230.0	114.0	13.8
Sample	Integer sample sort	32M 32-bit keys	24.7	13.2	13.0
Barnes	Hierarchical N-Body	1 Million Bodies	77.9	43.2	52.8
P-Ray	Ray Tracer	1 Million pixel image	23.5	17.9	156.2
Connect	Protocol Verification	SCI protocol, 2 proc	67.7	35.3	183.5
MurPHI	Connected Components		2.3	1.2	212.6
NOW-sort	Disk-to-Disk Sort	32M 100-byte records	127.2	56.9	817.4
Radb	Bulk version Radix	16M 32-bit keys	7.0	3.7	852.7

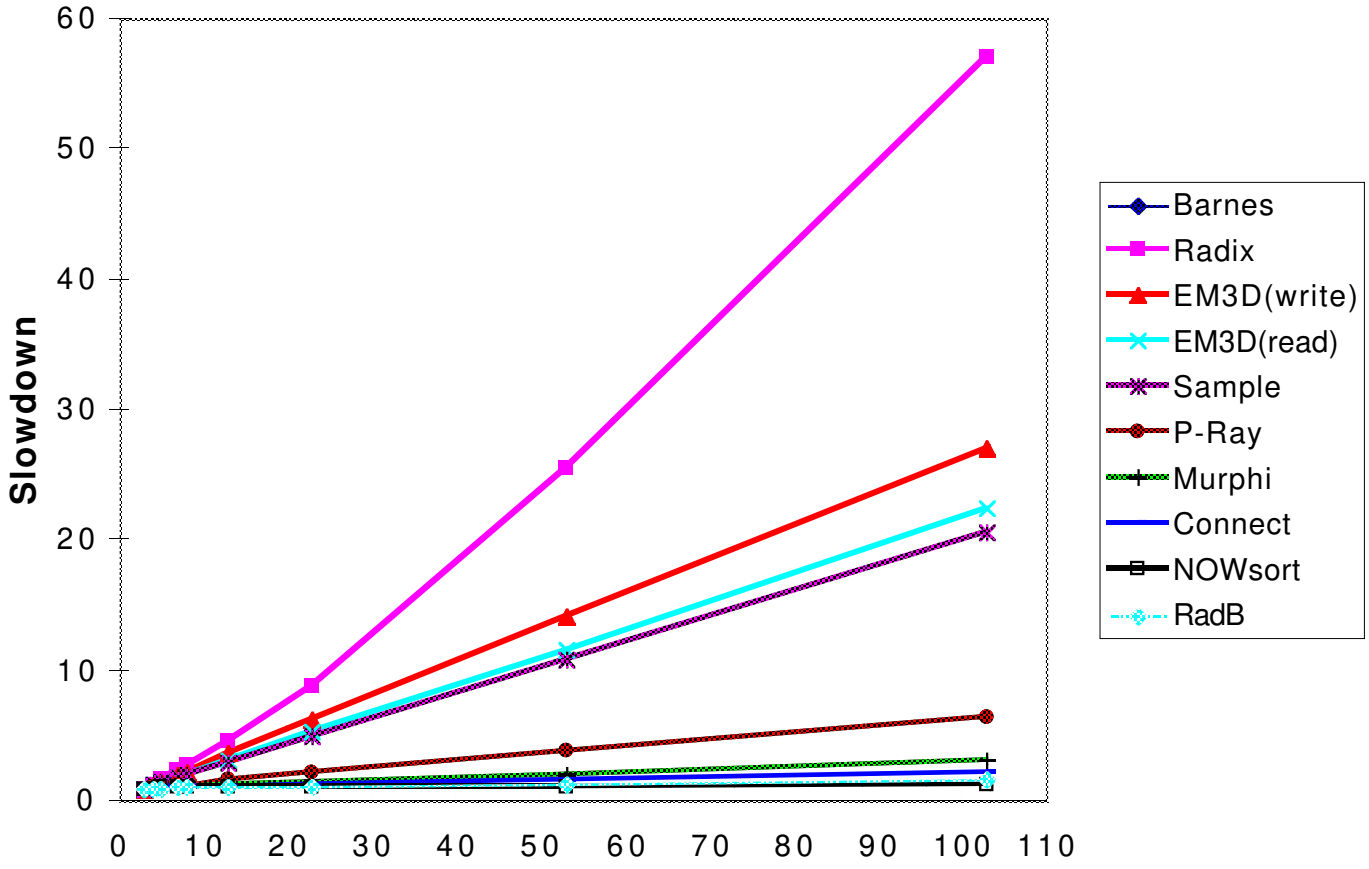
Baseline Communication

Program	µs/msg	Barrier	Bulks	Reads bulk	KB/s
Radix	6.1	408	0.0%	0.0%	26.7
EM3D(write)	8.0	122	0.0%	0.0%	0.6
EM3D(read)	13.8	369	0.0%	97.1%	0.0
Sample	13.0	1,203	0.0%	0.0%	0.0
Barnes	52.8	279	23.3%	20.6%	110.4
P-Ray	156.2	1,120	47.9%	96.5%	358.5
Connect	183.5	47	0.1%	67.4%	0.0
MurPHI	212.6	11,778	50.0%	0.0%	3876.6
NOW-sort	817.4	1,834	49.8%	0.0%	3125.1
Radb	852.7	25	34.7%	0.0%	33.6

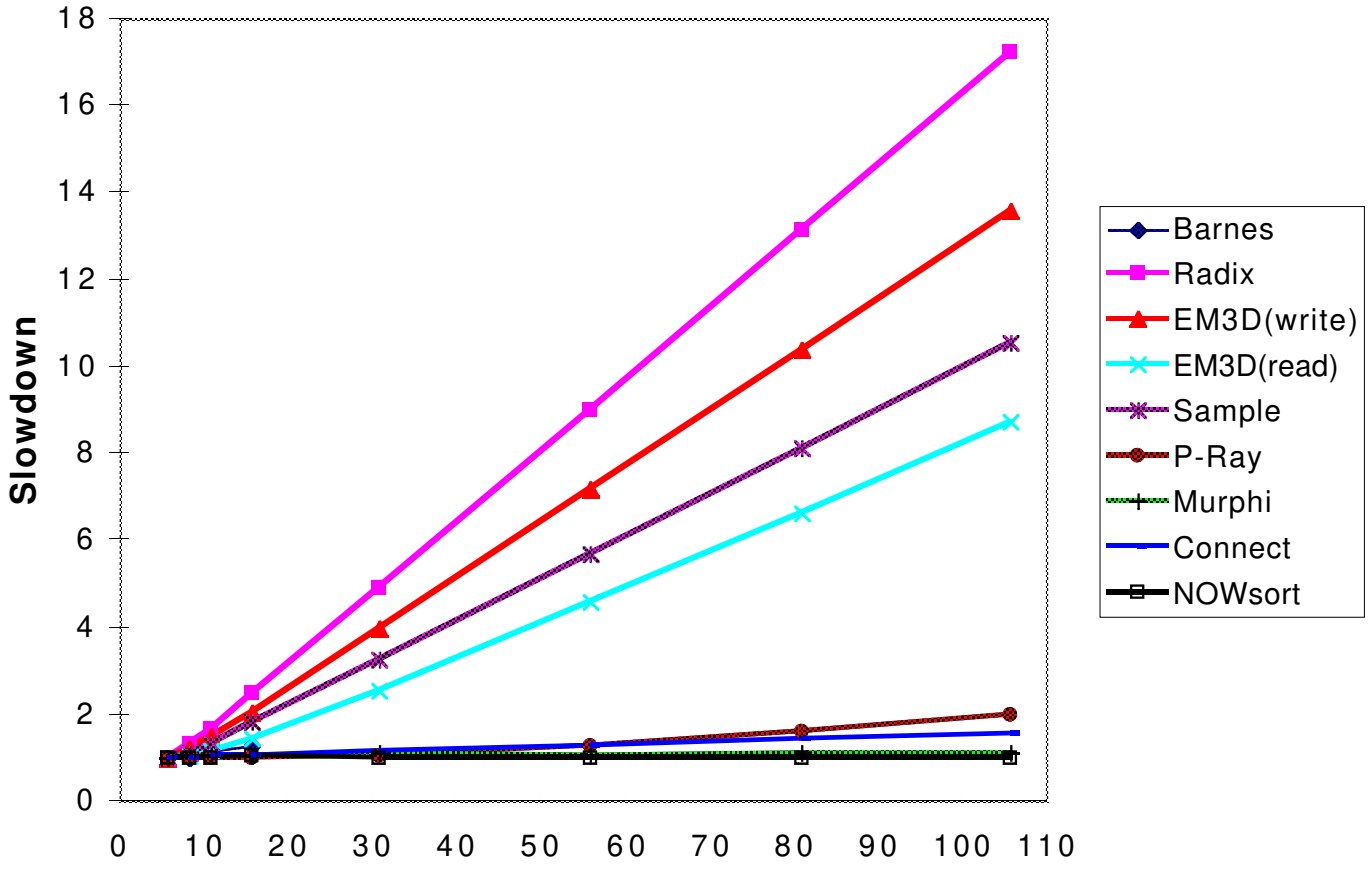
Communication Load Balance

Application Sensitivity to Communication Performance

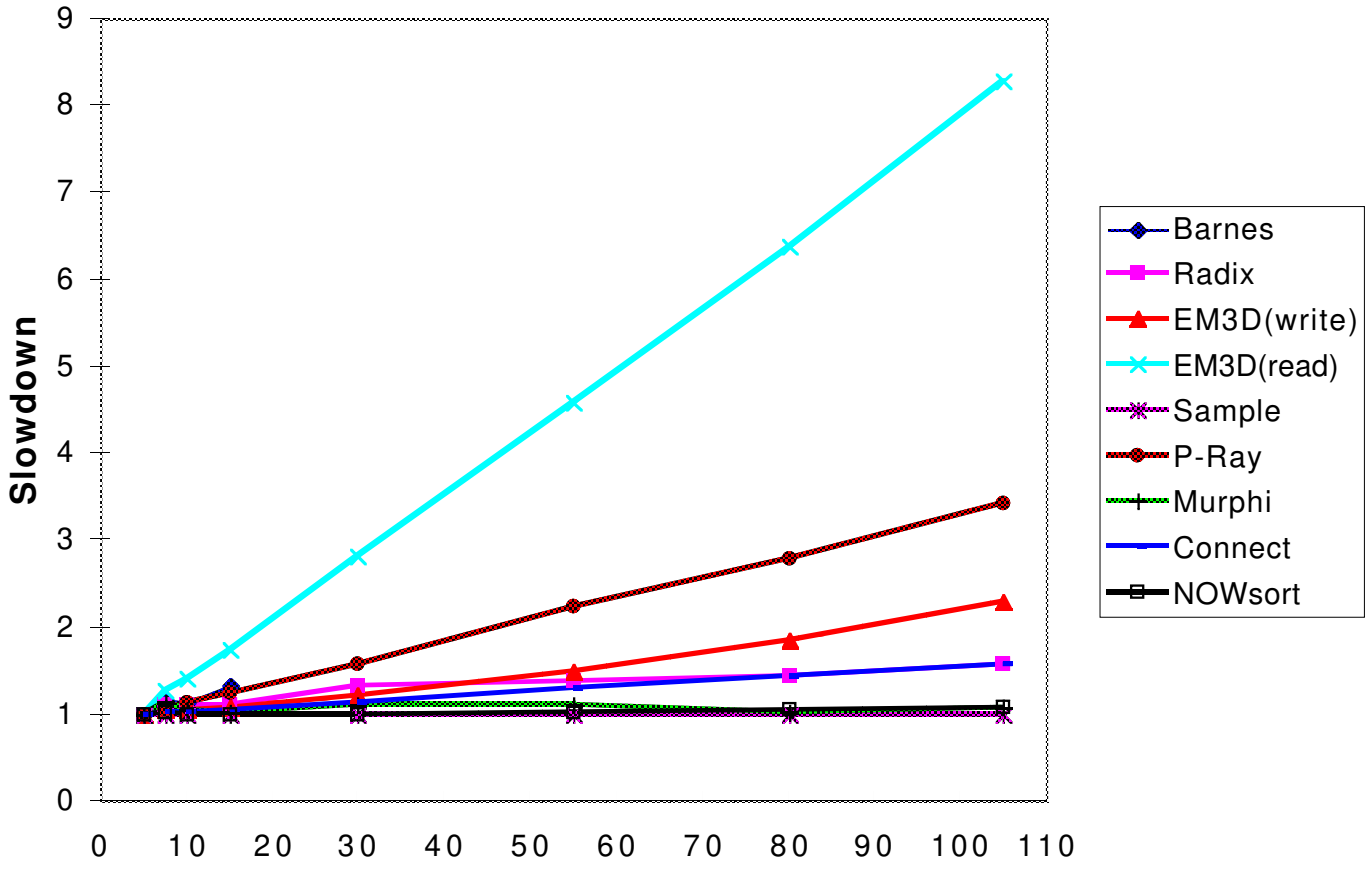
Sensitivity to Overhead (32 proc)



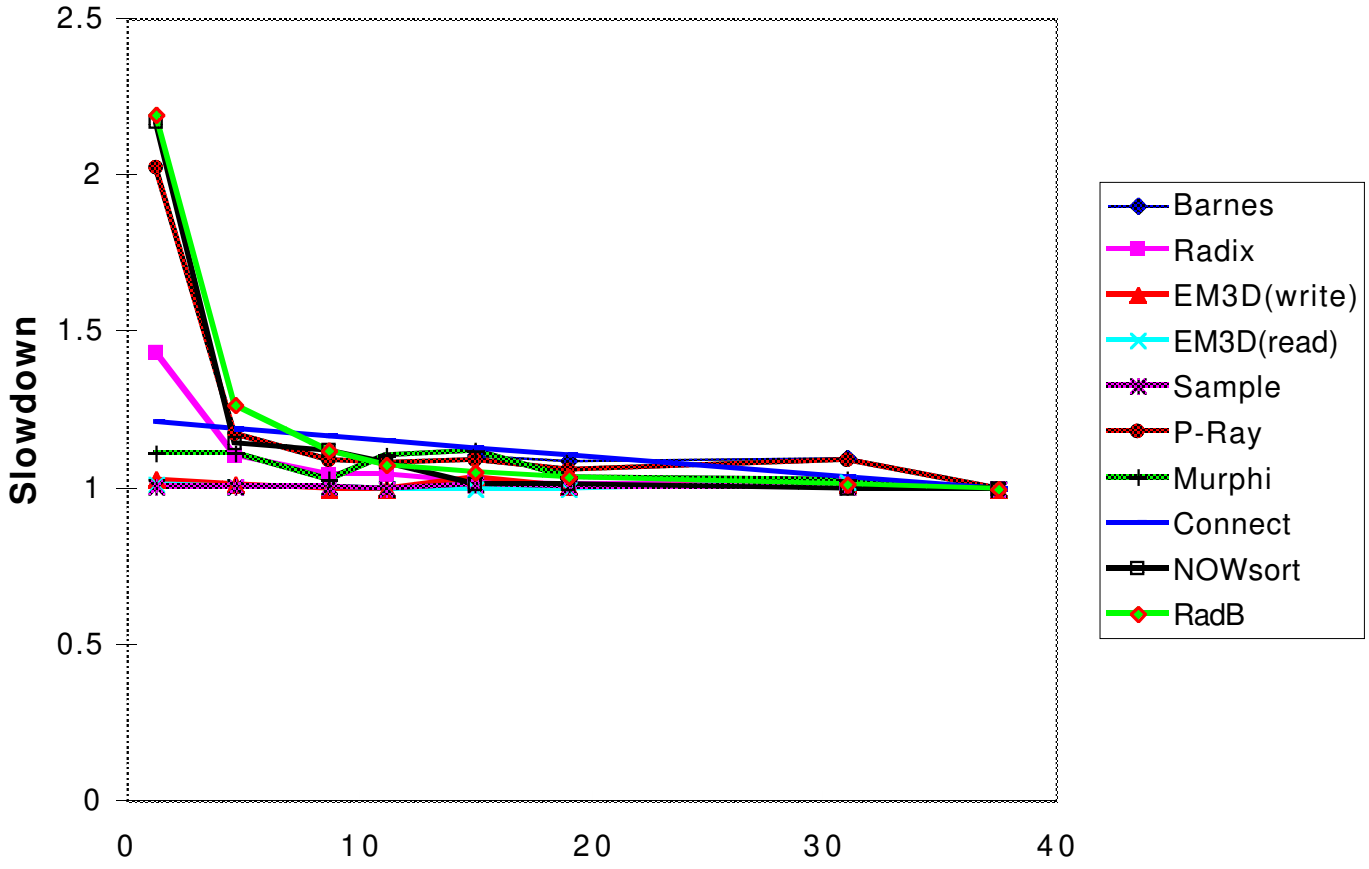
Sensitivity to gap (1/msg rate)



Sensitivity to Latency



Sensitivity to bulk BW (1/G)



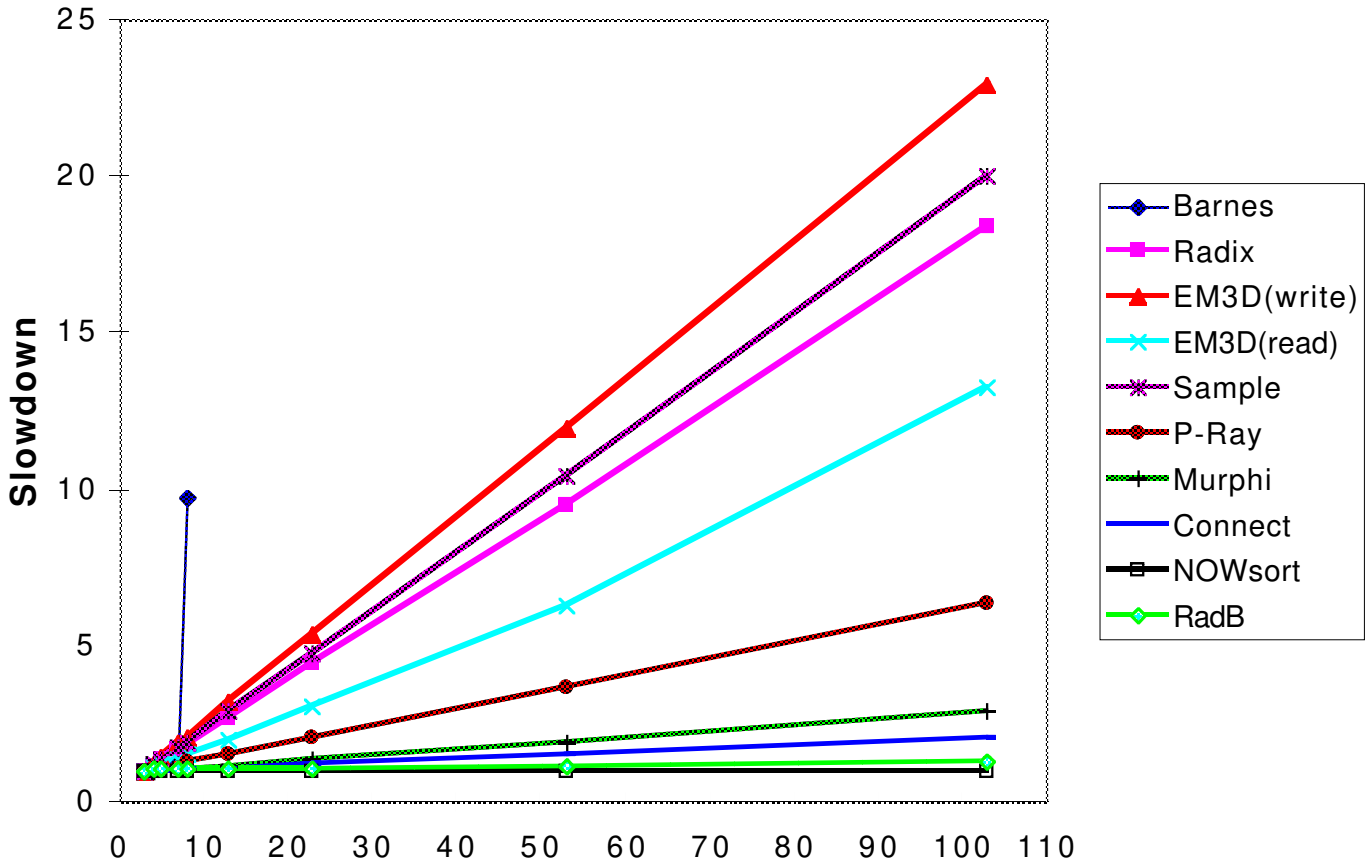
Modeling Effects of Overhead

- $T_{pred} = T_{orig} + 2 \times max\#msgs \times \Delta o$
  - request / response
  - proc with most msgs limits overall time

Overhead	Barnes	Radix	EM3Dw	EM3Dr	Sample	P-Ray	Murphi	Connect	NOWsort	RadB
2.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3.9	0.90	0.98	0.99	0.94	0.98	0.97	0.96	1.00	1.00	0.99
4.9	0.92	0.97	0.98	0.91	0.98	0.97	0.95	0.99	0.93	0.99
6.9	0.79	0.96	0.98	0.86	0.99	0.91	0.88	0.98	0.99	0.99
7.9		0.96	0.98	0.85	0.99	1.00	0.88	0.98	0.99	0.99
12.9		0.92	0.96	0.79	0.99	0.83	0.84	0.96	1.00	0.98
22.9		0.86	0.97	0.74	1.00	0.75	0.82	0.93	1.02	0.96
52.9		0.68	0.96	0.71	1.00	0.66	0.72	0.87	1.04	0.88
102.9		0.59	0.96	0.70	1.00	0.65	0.64	0.82	1.00	0.76

- **Why does this model under-predict?**

Sensitivity to Overhead (16 Proc)



Modeling Effects of gap

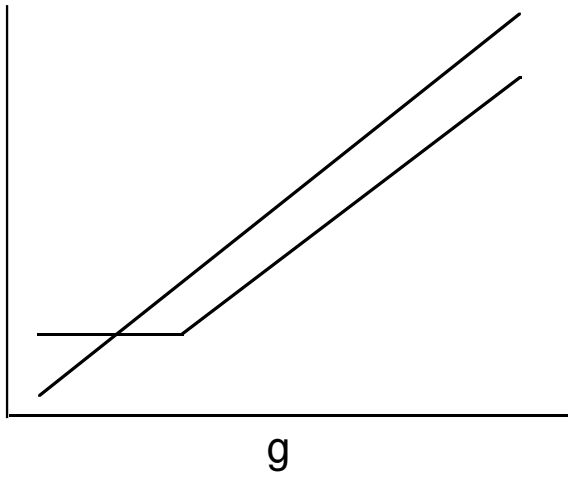
- **Uniform communication model**

$T_{pred} = T_{orig}$, **if g <** I. **average msg interval**

$= T_{orig} + m\,(g - I)$, **otherwise**

- **Bursty Communication**

$T_{pred} = T_{orig} + m\,\Delta g$

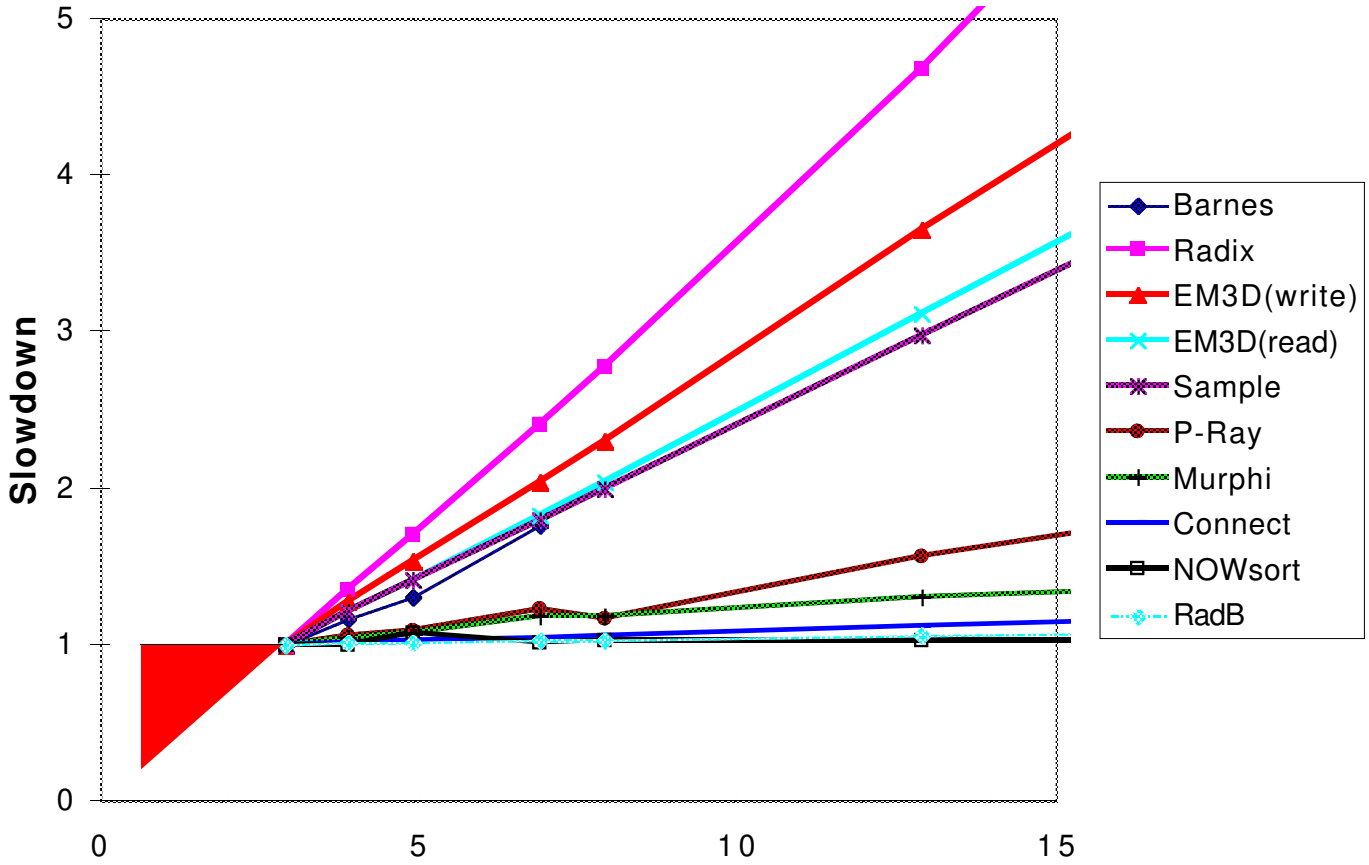


Evaluating BW Models

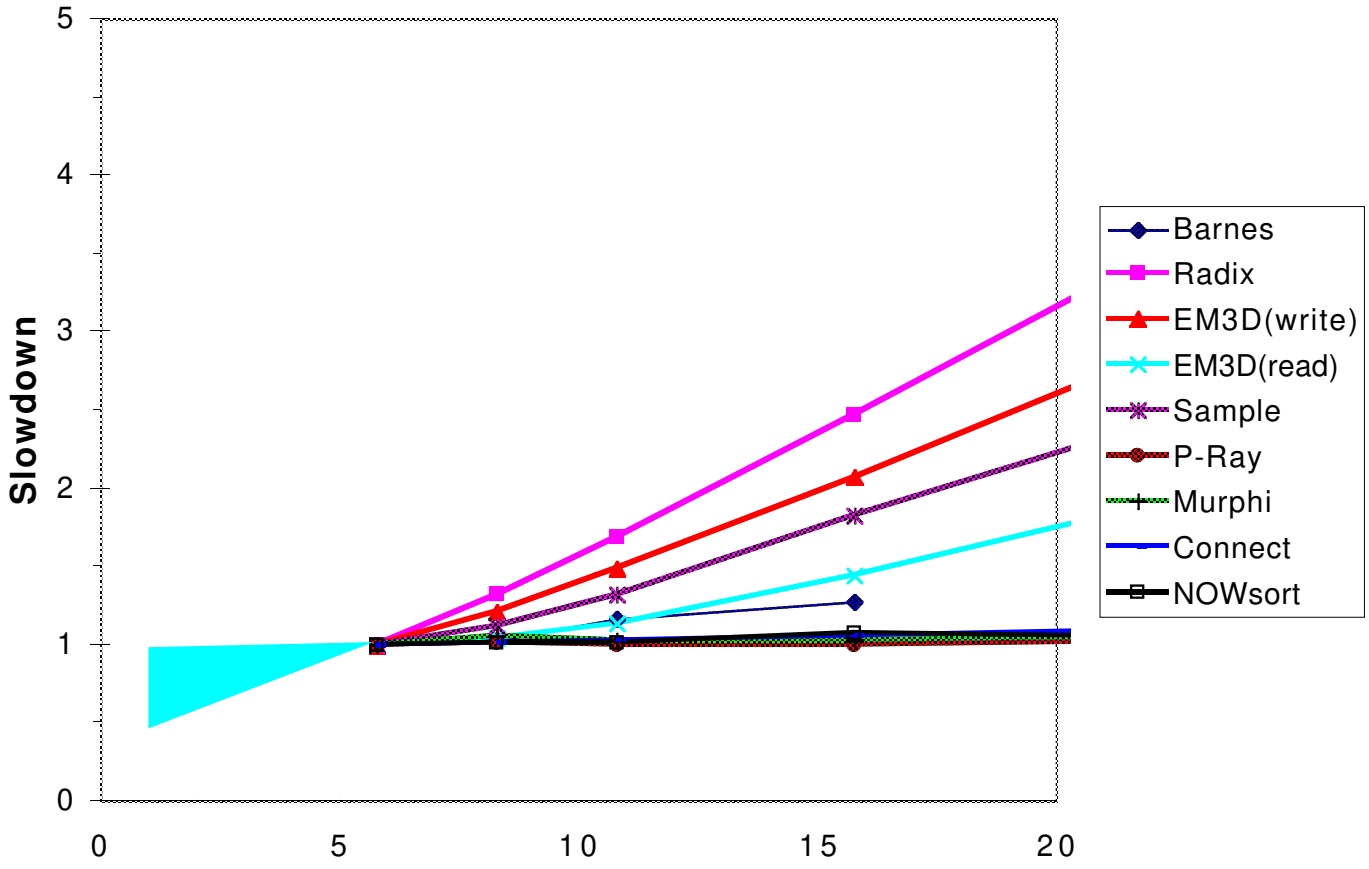
- Do not see the behavior of “uniform” model in any of the applications
- Pure bursty model gives good prediction
  - slightly over estimates in moderate cases

Do we need to push further?

Extrapolating to Low Overhead



Extrapolating to high msg rate



Conclusions

- Effort to improve communication performance is well rewarded
- Most sensitive to overhead
  - "does not parallelize well"
  - exascerbates serialization
- Surprisingly tolerant to latency
- Bulk BW not very critical
- Need to keep working on overhead!

- Valuable to use real experimental systems to explore relative to a model framework