Diploma Thesis: Communication Modeling and Placement of Parallel Applications

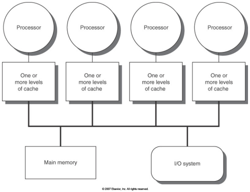
Supervisors: Nektarios Kozyris (CSLab/ECE/NTUA), Georgios Goumas (CSLAB/ECE/NTUA)

Georgios Christodoulis

ECE–NTUA

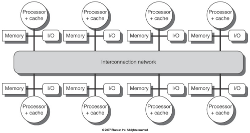
gchristodoulis@gmail.com

Parallel Architectures



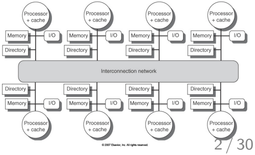
Shared Memory

1. Processors share the same physical memory
2. Local cache memory hierarchy
3. Connection via memory bus



Distributed Memory

1. Every processor has its own memory hierarcy
2. Processor communication via interconnection network



Hybrid Architectures

Dominant architecture for modern supercomputers.

Parallel Programming Models

OpenMP - Shared Memory Model

1. Compiler directives based tools
2. Suitable for shared memory architectures
3. Popular Schemas: Fork/Join, SPMD, parallel for, Master/Workers

MPI - Distributed Memory Model

1. Message Passing Library
2. SPMD - Every process runs the same program
3. Collective vs P2P
4. Blocking vs non-Blocking

OpenMP/MPI-Hybrid Model

1. Best fit for Hybrid architectures
2. OpenMP - intranode communication
3. MPI - communication through the interconnection network

Our System

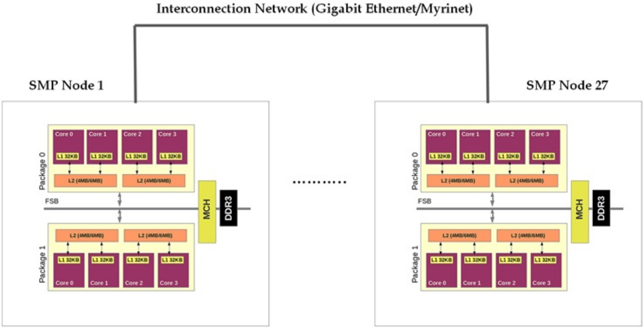
Architecture

- Hybrid Architecture
- Clovertown Architecture: Intel Xeon 2 GHz,

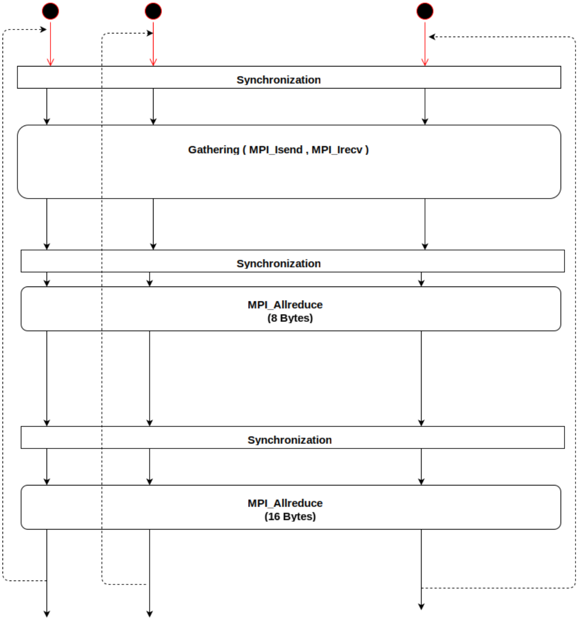
  32Kb L1 Cache/Core, 6MB L2 cache/package
- 4 Cores/package, 2 packages/node

Programming Model

MPI- Message Passing Interface



CGs Communication Pattern



Benchmarks - osu suite

P2P

- osu_latency ping-pong message exchange, blocking
- osu_multi_lat many pairs run simultaneously osu_latency
- osu_bw Sender sends back to back messages and waits for ack, non blocking
- osu_bibw Similar with osu_bw, both nodes send messages

Collective

osu_allreduce the benchmark when run from N processes, measures the min, max and average latency of the MPI_Allreduce operation

Contention on Switch

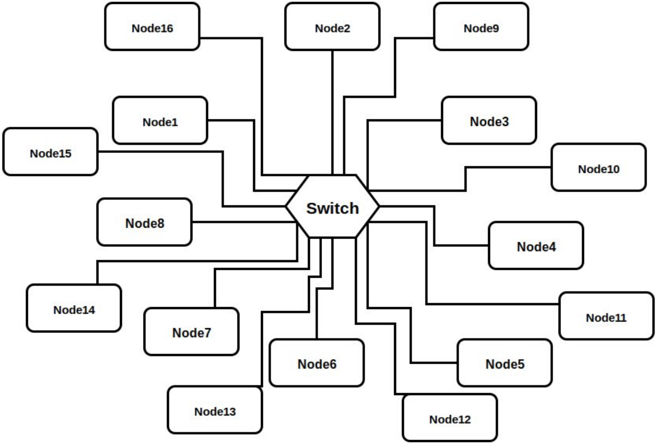
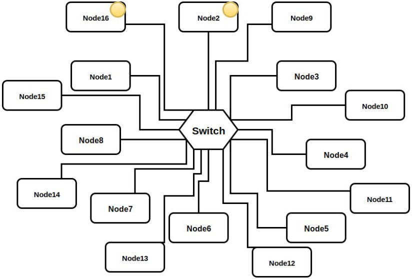
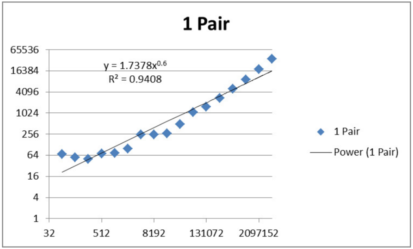


Figure: Systems instance for contention on switch testing

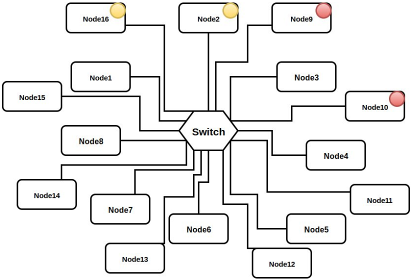
Contention on Switch

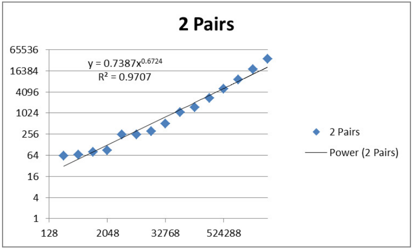




$$C_2^1(s) = \begin{cases} 50\mu sec & \text{, if size } s < 64 \text{ Bytes} \\ 1.7378 \times s^{0.6}\mu sec & \text{, if size } s \geq 64 \text{ Bytes} \end{cases}$$

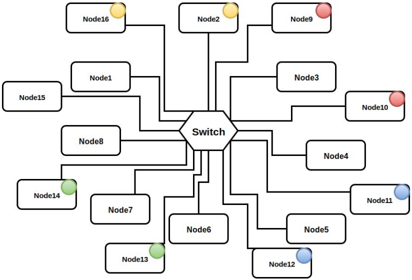
Contention on Switch

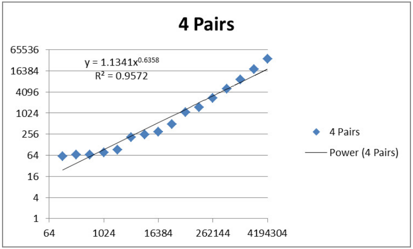




$$C_2^1(s) = \begin{cases} 50\mu sec & \text{, if size } s < 256 \text{ Bytes} \\ 0.7387 \times s^{0.6724} \mu sec & \text{, if size } s \geq 256 \text{ Bytes} \end{cases}$$

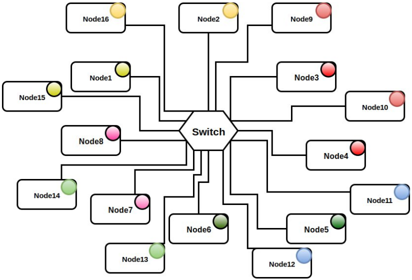
Contention on Switch

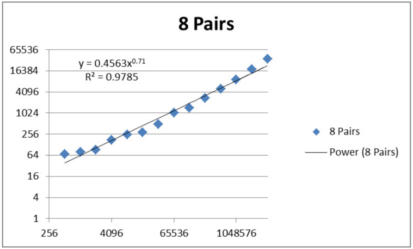




$$C_4^1(s) = \begin{cases} 50\mu sec & \text{, if size } s < 256 \text{ Bytes} \\ 1.1341 \times s^{0.6358}\mu sec & \text{, if size } s \geq 256 \text{ Bytes} \end{cases}$$

Contention on Switch





$$C_8^1(s) = \begin{cases} 50\mu sec & \text{, if size } s < 256 \text{ Bytes} \\ 0.4563 \times s^{0.71} \mu sec & \text{, if size } s \geq 256 \text{ Bytes} \end{cases}$$

Contention on Switch

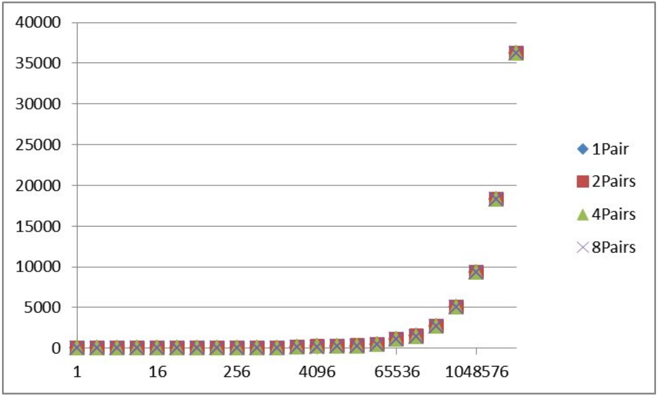


Figure: Complete indipendancy on switch access.

Contention on NIC

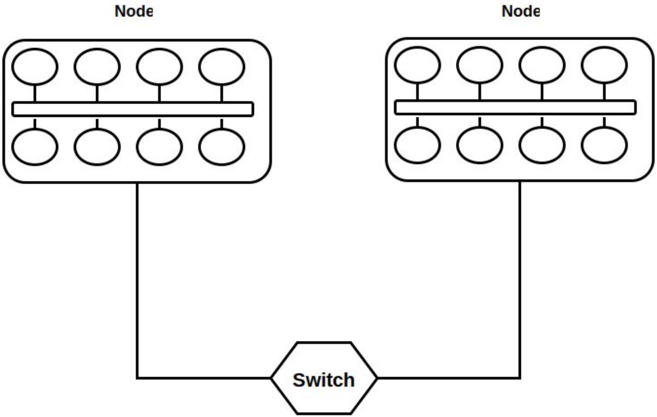
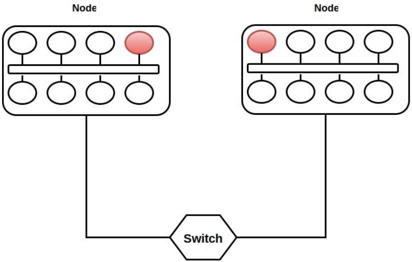
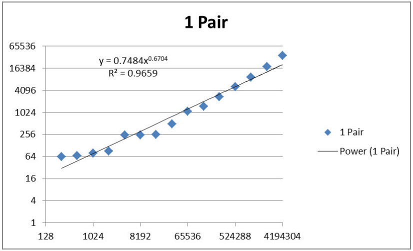


Figure: Instance of system for contention testing

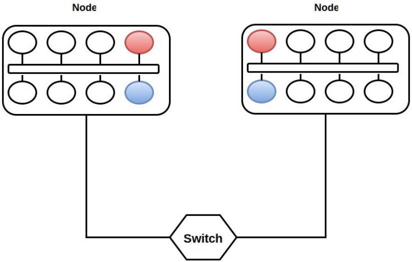
Contention on NIC

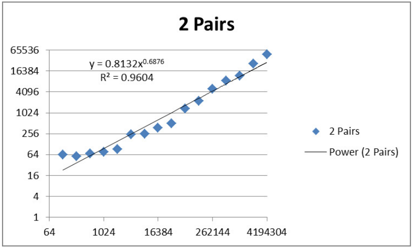




$$C_2^1(s) = \begin{cases} 50\mu sec & \text{, if size } s < 256 \text{ Bytes} \\ 0.7484 \times s^{0.6704}\mu sec & \text{, if size } s \geq 256 \text{ Bytes} \end{cases}$$

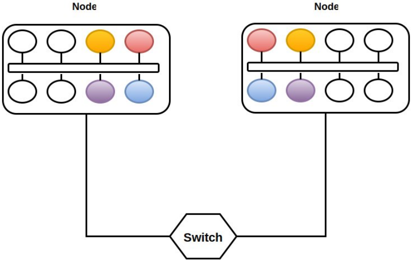
Contention on NIC

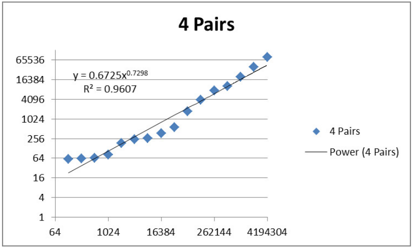




$$C_2^2(s) = \begin{cases} 50\mu sec & \text{, if size } s < 64 \text{ Bytes} \\ 0.8132 \times s^{0.6876}\mu sec & \text{, if size } s \geq 64 \text{ Bytes} \end{cases}$$

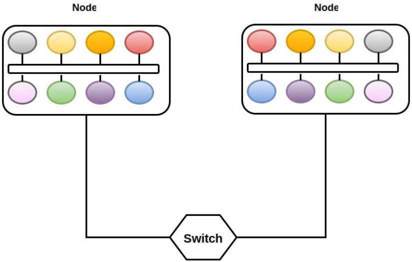
Contention on NIC

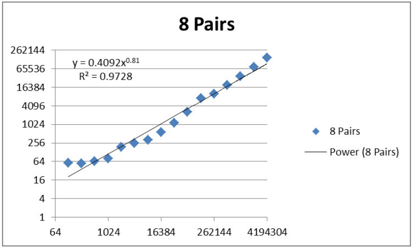




$$C_2^4(s) = \begin{cases} 50\mu sec & \text{, if size } s < 64 \text{ Bytes} \\ 0.6725 \times s^{0.7298} \mu sec & \text{, if size } s \geq 64 \text{ Bytes} \end{cases}$$

Contention on NIC





$$C_2^8(s) = \begin{cases} 50\mu sec & \text{, if size } s < 64 \text{ Bytes} \\ 0.4092 \times s^{0.81} \mu sec & \text{, if size } s \geq 64 \text{ Bytes} \end{cases}$$

Contention on NIC

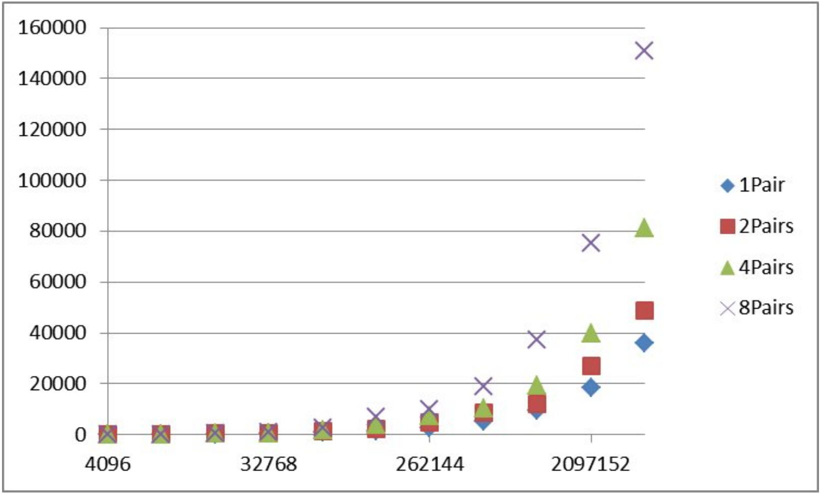


Figure: Appearance of Contention on NIC

Intranode Effect

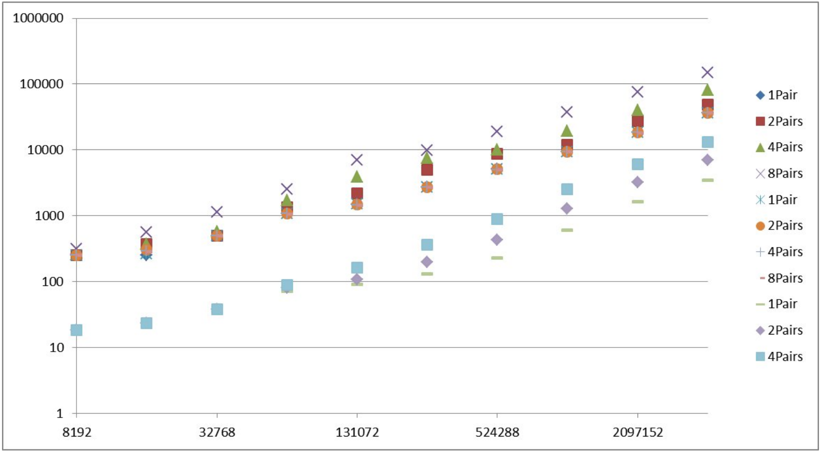


Figure: Communication Overview

Useful Metrics

- Maximum Send Latency
- Maximum Receive Latency
- Sum of Send Latencies
- Sum of Receive Latencies
- S+R

Results - Gathering Session

Table: Prediction Results for Gathering Session

Matrix	$S + R$	Actual Gathering Time	Relative Deviation
af_5_k101	0.49985	0.49822	0.0032
af_shell10	0.35531	0.29756	0.1940
af_shell9	0.47806	0.40173	0.1900
apache2	0.42745	0.35508	0.20381
bmw3_2	0.34752	0.34752	0 (HIT)
bmwcra_1	0.93015	0.93015	0 (HIT)
bone010	0.49285	0.49285	0 (HIT)
boneS10	0.42705	0.42705	0 (HIT)
crankseg_2	0.45069	0.45069	0 (HIT)
F1	0.38515	0.38515	0 (HIT)

Results - Gathering Session

Table: Prediction Results for Gathering Session

Matrix	$S + R$	Actual Gathering Time	Relative Deviation
G3_circuit	2.44527	2.44527	0 (HIT)
Ga41As41H72	0.22786	0.22786	0 (HIT)
helm2d03	0.26700	0.26700	0 (HIT)
hood	0.51297	0.39284	0.3057
inline_1	0.89312	0.89312	0 (HIT)
kkt_power	0.42427	0.41462	0.0232
ldoor	0.32107	0.32107	0 (HIT)
msdoor	0.88584	0.88584	0 (HIT)
nd12k	1.14328	1.14328	0 (HIT)

Results - Gathering Session

Table: Prediction Results for Gathering Session

Matrix	$S + R$	Actual Gathering Time	Relative Deviation
nd24k	0.72550	0.72550	0 (HIT)
nd6k	0.24027	0.22966	0.0461
parabolic_fem	0.23146	0.23146	0 (HIT)
pwtk	0.28839	0.24462	0.1789
s3dkq4m2	0.19154	0.19154	0 (HIT)
ship_001	2.13303	2.13303	0 (HIT)
Si41Ge41H72	2.48232	2.48232	0 (HIT)
Si87H76	0.23744	0.23744	0 (HIT)
thermal2	0.47869	0.47869	0 (HIT)
thread	0.31130	0.31130	0 (HIT)

Collective Communication

Table: Collective Communication Latency, Message size $<$ 64B

Latency(μsec), 2 Nodes		
2 Pairs	4 Pairs	8 Pairs
84.79	121.48	171.23

Table: Collective Communication Latency, Message size $<$ 64B

Latency(μsec), 4 Nodes			
2 Pairs	4 Pairs	8 Pairs	16 Pairs
120.22	144.21	185.33	255.66

Table: Collective Communication Latency, Message size $<$ 64B

Latency(μsec), 8 Nodes				
2 Pairs	4 Pairs	8 Pairs	16 Pairs	32 Pairs
104.01	191.17	219.28	242.27	346.17

Collective Communication Predictor

Table: Collective Communication Predictor for CG

	Latency(μsec)		
Processes Per Node	2 Nodes	4 Nodes	8 Nodes
4 PPN	242.96	370.66	484.54
8 PPN	342.46	511.32	692.34

Results - Entire Communication

Table: Prediction Results for CG

Matrix	$S + R + A$	Actual Communication Time	Relative Deviation
af_5_k101	0.92685	0.92685	0 (HIT)
af_shell10	0.75131	0.72214	0.0011
af_shell9	0.87909	0.87909	0 (HIT)
apache2	0.81645	0.72729	0.1225
bmw3_2	0.77252	0.77252	0 (HIT)
bmwcra_1	1.34415	1.34415	0 (HIT)
bone010	0.90385	0.90385	0 (HIT)
boneS10	0.82305	0.82305	0 (HIT)
crankseg_2	0.84369	0.84369	0 (HIT)
F1	0.82350	0.78090	0.0545

Results - Entire Communication

Table: Prediction Results for CG

Matrix	$S + R + A$	Actual Communication Time	Relative Deviation
G3_circuit	2.84727	2.84727	0 (HIT)
Ga41As41H72	0.61586	0.61586	0 (HIT)
helm2d03	0.66215	0.66215	0 (HIT)
hood	1.13997	0.79384	0.4360
inline_1	1.33412	1.33412	0 (HIT)
kkt_power	0.84827	0.77354	0.0966
ldoor	0.71707	0.69761	0.0278
msdoor	1.28484	1.28484	0 (HIT)
nd12k	1.55428	1.55428	0 (HIT)

Results - Entire Communication

Table: Prediction Results for CG

Matrix	$S + R + A$	Actual Communication Time	Relative Deviation
nd24k	1.11750	1.01191	0.1043
nd6k	0.50727	0.50727	0 (HIT)
parabolic_fem	0.62846	0.62846	0 (HIT)
pwtk	0.67739	0.67739	0 (HIT)
s3dkq4m2	0.58054	0.58054	0 (HIT)
ship_001	2.52803	2.52803	0 (HIT)
Si41Ge41H72	2.87932	2.87932	0 (HIT)
Si87H76	0.65244	0.65244	0 (HIT)
thermal2	0.86769	0.86769	0 (HIT)
thread	0.72289	0.72289	0 (HIT)

Results

Accuracy on Gathering Session

$$\frac{hit}{miss} = \frac{21}{8}$$

$$\text{Deviation}_{S+R} = 4.05\%$$

Accuracy on the entire communication pattern

$$\frac{hit}{miss} = \frac{22}{7}$$

$$\text{Deviation}_{S+R+A} = 2.91\%$$

Conclusions

- We attempted to predict the optimal placement for CG
- We performed a series of benchmarks in order to explore systems behavior
- We finally predicted the optimal placement for both the dominant session as well as the entire communication of CG, with satisfying accuracy and deviation.