RWTH AACHEN UNIVERSITY

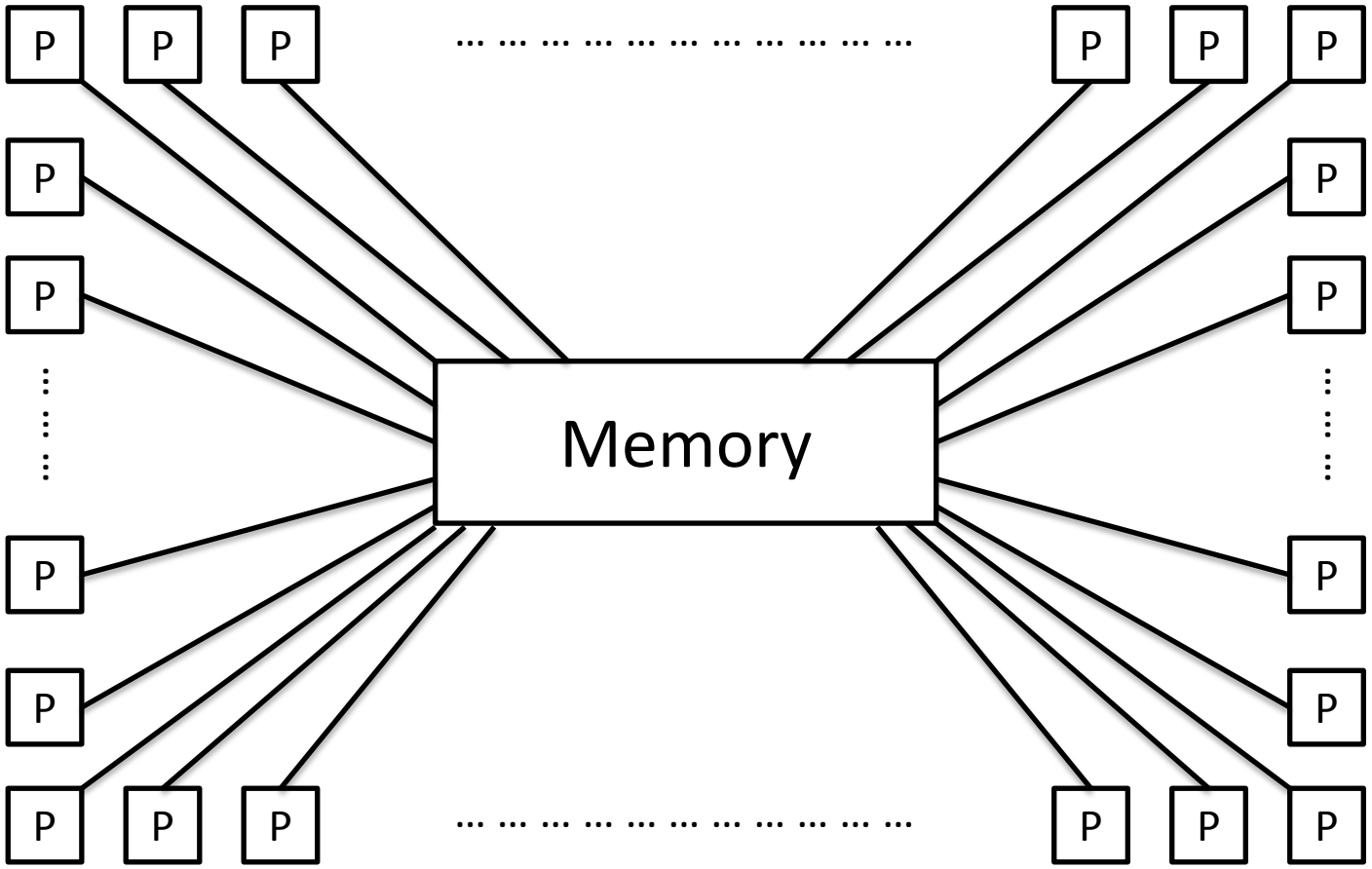
Towards NUMA Support with Distance Information

Dirk Schmidl, Christian Terboven, Dieter an Mey
{schmidl | terboven | anmey}@rz.rwth-aachen.de

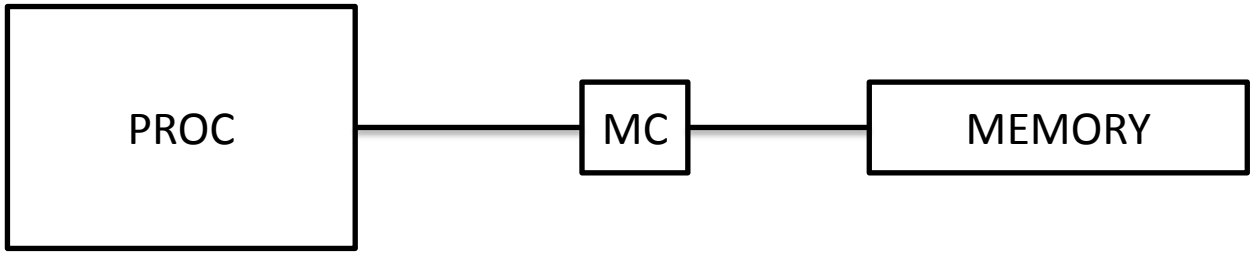
Agenda

- Topology of modern Hardware
- Distance Matrix
- Usage of Distance Information
- Example – Array Reduction

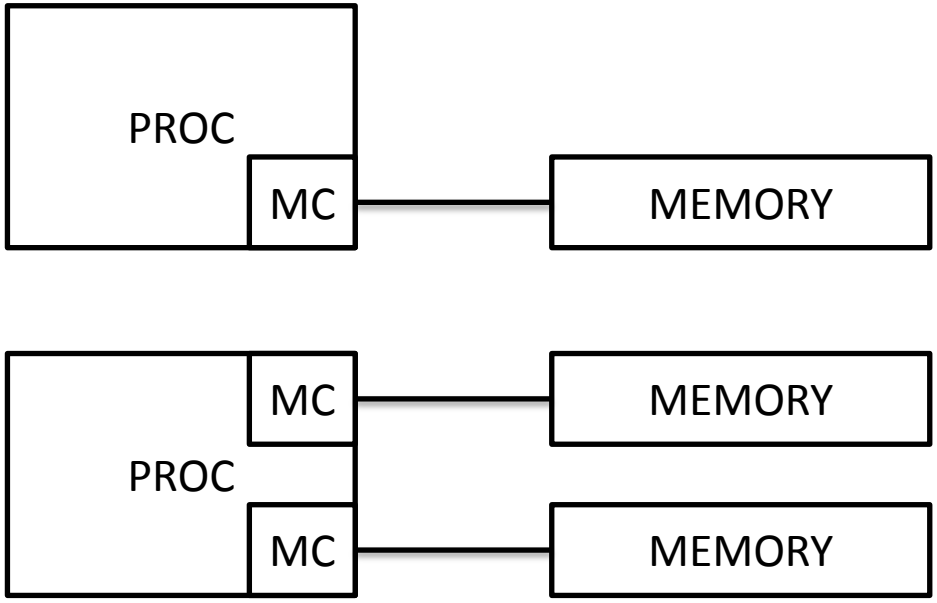
Shared Memory in OpenMP



External Memory Controller

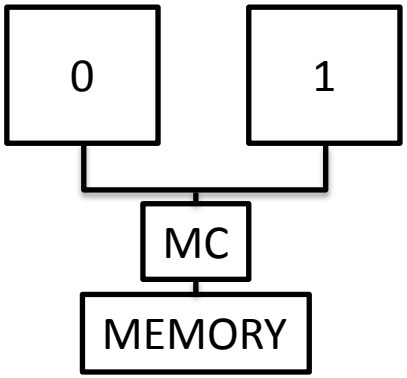


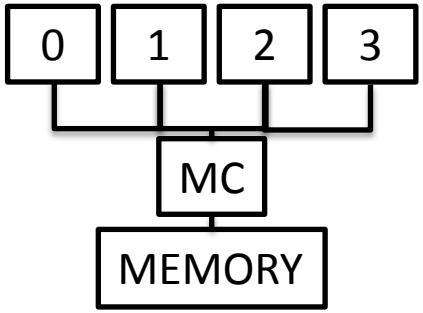
Memory Controller on the chip



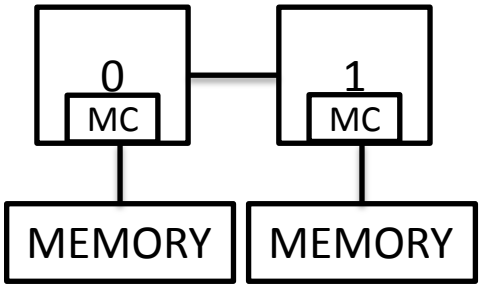
Shared Memory Systems - Board

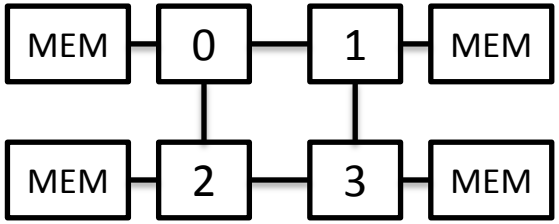
Uniform Memory Access

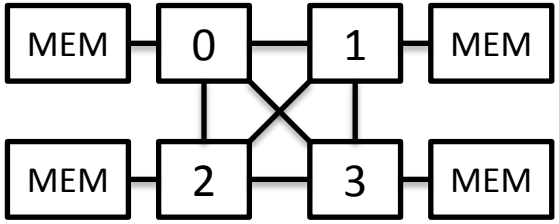




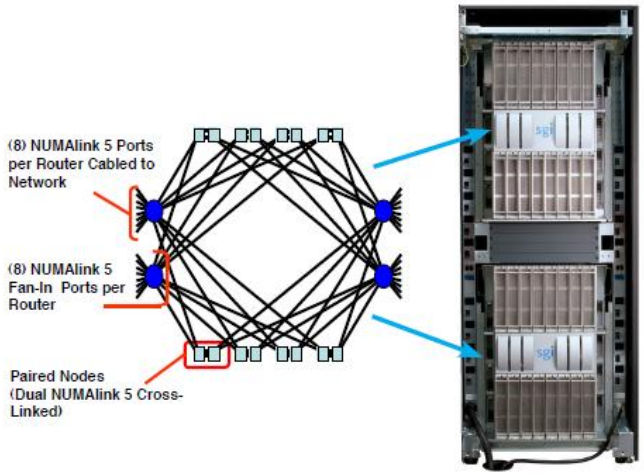
Non Uniform Memory Access





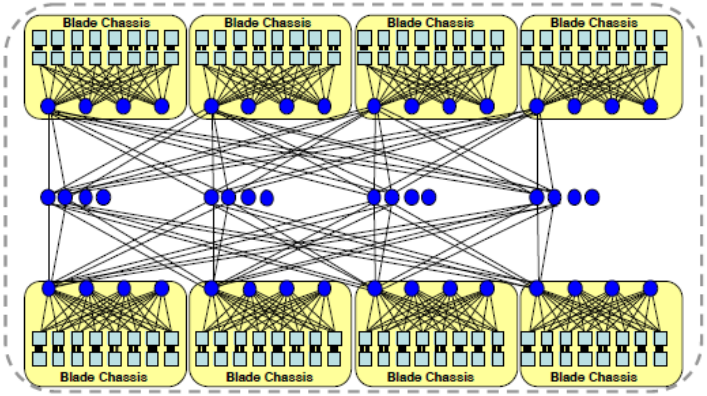


Shared Memory Systems – Several Board



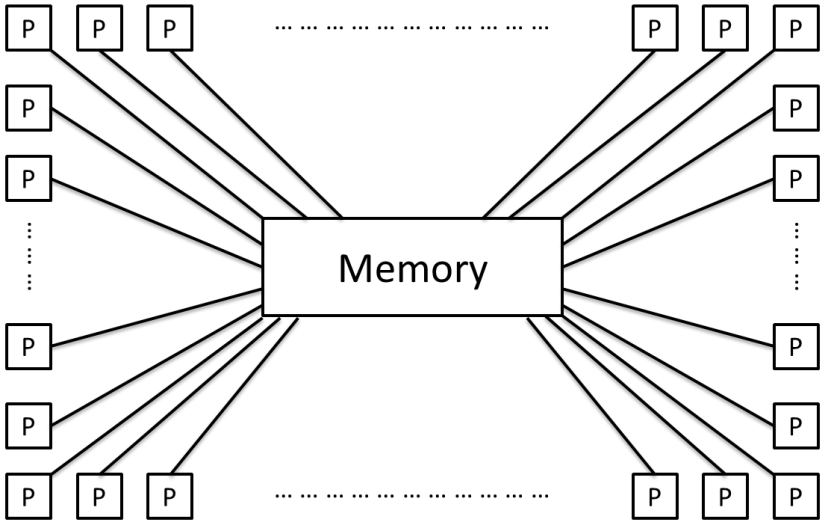
Switched dual-plane topology inside of one blade chassis

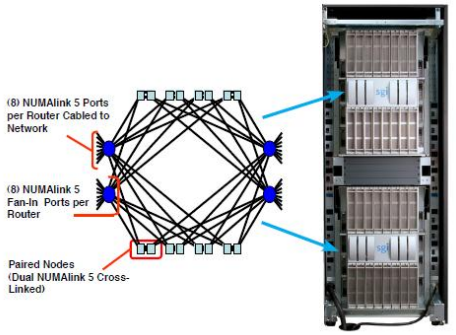
SGI Altix UV
- up to 2048 cores
- up to 4096 Hyperthreads

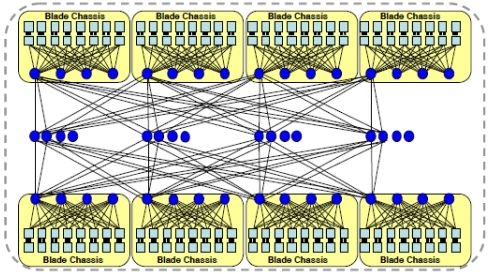


16 blade chassis can be combined with a fat-tree topology

Pictures from: *Technical Advances in the SGI Altix UV* Architecture
Available at: http://www.sgi.com/pdfs/4192.pdf



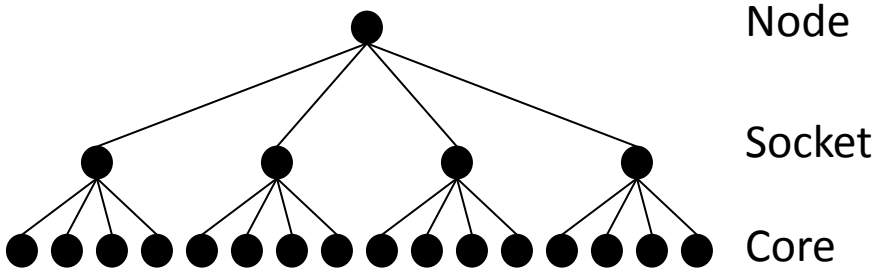


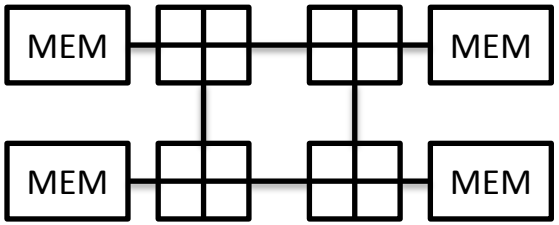


The OpenMP Memory Model does not reflect current hardware.

Description of the Topology

Tree Based Description





Matrix Description

Core	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3
1	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3
2	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3
3	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3
4	2	2	2	2	1	1	1	1	3	3	3	3	2	2	2	2
5	2	2	2	2	1	1	1	1	3	3	3	3	2	2	2	2
6	2	2	2	2	1	1	1	1	3	3	3	3	2	2	2	2
7	2	2	2	2	1	1	1	1	3	3	3	3	2	2	2	2
8	2	2	2	2	3	3	3	3	1	1	1	1	2	2	2	2
9	2	2	2	2	3	3	3	3	1	1	1	1	2	2	2	2
10	2	2	2	2	3	3	3	3	1	1	1	1	2	2	2	2
11	2	2	2	2	3	3	3	3	1	1	1	1	2	2	2	2
12	3	3	3	3	2	2	2	2	2	2	2	2	1	1	1	1
13	3	3	3	3	2	2	2	2	2	2	2	2	1	1	1	1
14	3	3	3	3	2	2	2	2	2	2	2	2	1	1	1	1
15	3	3	3	3	2	2	2	2	2	2	2	2	1	1	1	1

System Locality Distance Information Table

- Distance Information from the BIOS
- Distance between NUMA nodes
- Values are vendor specific

Socket	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
0	10	13	40	40	40	40	48	48	48	48	55	55	55	55	62	62	62	62	69	69	55	55	62	62	48	48	55	55	40	40	48	48
1	13	10	40	40	40	40	48	48	48	48	55	55	55	55	62	62	62	62	69	69	55	55	62	62	48	48	55	55	40	40	48	48
2	40	40	10	13	48	48	40	40	55	55	48	48	62	62	55	55	69	69	62	62	62	62	55	55	55	55	48	48	48	48	40	40
3	40	40	13	10	48	48	40	40	55	55	48	48	62	62	55	55	69	69	62	62	62	62	55	55	55	55	48	48	48	48	40	40
4	40	40	48	48	10	13	40	40	40	40	48	48	48	48	55	55	55	55	62	62	62	62	69	69	55	55	62	62	48	48	55	55
5	40	40	48	48	13	10	40	40	40	40	48	48	48	48	55	55	55	55	62	62	62	62	69	69	55	55	62	62	48	48	55	55
6	48	48	40	40	40	40	10	13	48	48	40	40	55	55	48	48	62	62	55	55	69	69	62	62	62	62	55	55	55	55	48	48
7	48	48	40	40	40	40	13	10	48	48	40	40	55	55	48	48	62	62	55	55	69	69	62	62	62	62	55	55	55	55	48	48
8	48	48	55	55	40	40	48	48	10	13	40	40	40	40	48	48	48	48	55	55	55	55	62	62	62	62	69	69	55	55	62	62
9	48	48	55	55	40	40	48	48	13	10	40	40	40	40	48	48	48	48	55	55	55	55	62	62	62	62	69	69	55	55	62	62
10	55	55	48	48	48	48	40	40	40	40	10	13	48	48	40	40	55	55	48	48	62	62	55	55	69	69	62	62	62	62	55	55
11	55	55	48	48	48	48	40	40	40	40	13	10	48	48	40	40	55	55	48	48	62	62	55	55	69	69	62	62	62	62	55	55
12	55	55	62	62	48	48	55	55	40	40	48	48	10	13	40	40	40	40	48	48	48	48	55	55	55	55	62	62	62	62	69	69
13	55	55	62	62	48	48	55	55	40	40	48	48	13	10	40	40	40	40	48	48	48	48	55	55	55	55	62	62	62	62	69	69
14	62	62	55	55	55	55	48	48	48	48	40	40	40	40	10	13	48	48	40	40	55	55	48	48	62	62	55	55	69	69	62	62
15	62	62	55	55	55	55	48	48	48	48	40	40	40	40	13	10	48	48	40	40	55	55	48	48	62	62	55	55	69	69	62	62
16	62	62	69	69	55	55	62	62	48	48	55	55	40	40	48	48	10	13	40	40	40	40	48	48	48	48	55	55	55	55	62	62
17	62	62	69	69	55	55	62	62	48	48	55	55	40	40	48	48	13	10	40	40	40	40	48	48	48	48	55	55	55	55	62	62
18	69	69	62	62	62	62	55	55	55	55	48	48	48	48	40	40	40	40	10	13	48	48	40	40	55	55	48	48	62	62	55	55
19	69	69	62	62	62	62	55	55	55	55	48	48	48	48	40	40	40	40	13	10	48	48	40	40	55	55	48	48	62	62	55	55
20	55	55	62	62	62	62	69	69	55	55	62	62	48	48	55	55	40	40	48	48	10	13	40	40	40	40	48	48	48	48	55	55
21	55	55	62	62	62	62	69	69	55	55	62	62	48	48	55	55	40	40	48	48	13	10	40	40	40	40	48	48	48	48	55	55
22	62	62	55	55	69	69	62	62	62	62	55	55	55	55	48	48	48	48	40	40	40	40	10	13	48	48	40	40	55	55	48	48
23	62	62	55	55	69	69	62	62	62	62	55	55	55	55	48	48	48	48	40	40	40	40	13	10	48	48	40	40	55	55	48	48
24	48	48	55	55	55	55	62	62	62	62	69	69	55	55	62	62	48	48	55	55	40	40	48	48	10	13	40	40	40	40	48	48
25	48	48	55	55	55	55	62	62	62	62	69	69	55	55	62	62	48	48	55	55	40	40	48	48	13	10	40	40	40	40	48	48
26	55	55	48	48	62	62	55	55	69	69	62	62	62	62	55	55	55	55	48	48	48	48	40	40	40	40	10	13	48	48	40	40
27	55	55	48	48	62	62	55	55	69	69	62	62	62	62	55	55	55	55	48	48	48	48	40	40	40	40	13	10	48	48	40	40
28	40	40	48	48	48	48	55	55	55	55	62	62	62	62	69	69	55	55	62	62	48	48	55	55	40	40	48	48	10	13	40	40
29	40	40	48	48	48	48	55	55	55	55	62	62	62	62	69	69	55	55	62	62	48	48	55	55	40	40	48	48	13	10	40	40
30	48	48	40	40	55	55	48	48	62	62	55	55	69	69	62	62	62	62	55	55	55	55	48	48	48	48	40	40	40	40	10	13
31	48	48	40	40	55	55	48	48	62	62	55	55	69	69	62	62	62	62	55	55	55	55	48	48	48	48	40	40	40	40	13	10

SLIT for a SGI Altix UV

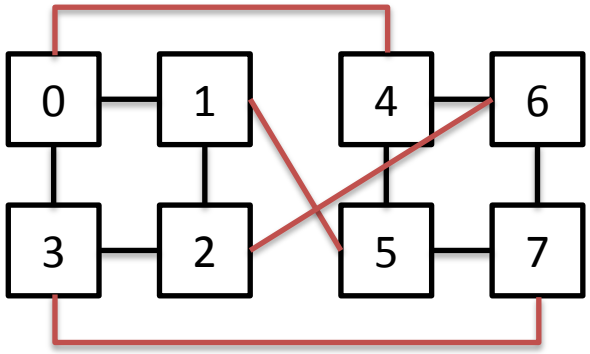
	0	1	2	3	4	5
0	10	13	40	40	40	40
1	13	10	40	40	40	40
2	40	40	10	13	48	48
3	40	40	13	10	48	48
4	40	40	48	48	10	13
5	40	40	48	48	13	10

System Locality Distance Information Table

- Information in the SLIT is vendor specific
- Sometimes there is less information than for the Altix UV
- The real Topology might not be reflected in the SLIT entries

Socket	0	1	2	3	4	5	6	7
0	10	12	12	12	12	12	12	12
1	12	10	12	12	12	12	12	12
2	12	12	10	12	12	12	12	12
3	12	12	12	10	12	12	12	12
4	12	12	12	12	10	12	12	12
5	12	12	12	12	12	10	12	12
6	12	12	12	12	12	12	10	12
7	12	12	12	12	12	12	12	10

SLIT for an 8 Socket Intel Nehalem



Topology of the 8 Socket Intel Nehalem

Socket	0	1	2	3	4	5	6	7
0	0	1	2	1	1	2	2	2
1	1	0	1	2	2	1	2	2
2	2	1	0	1	2	2	1	2
3	1	2	1	0	2	2	2	1
4	1	2	2	2	0	1	1	2
5	2	1	2	2	1	0	2	1
6	2	2	1	2	1	2	0	1
7	2	2	2	1	2	1	1	0

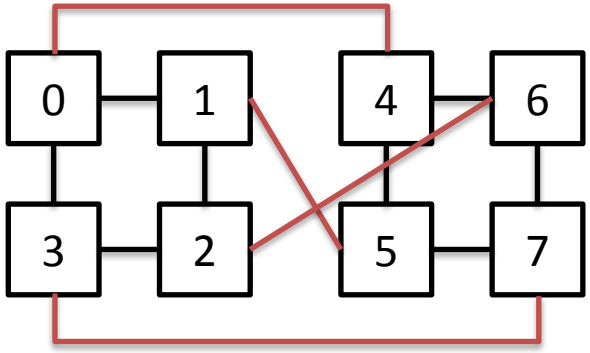
Number of Hops between NUMA nodes

Measuring Distance Information

Measuring Aspects:

- Bandwidth <-> Latency
- 1 Thread <-> Many Threads
- idle system <-> busy system

e.g. on an idle system the bandwidth differences between local and remote memory were < 5%



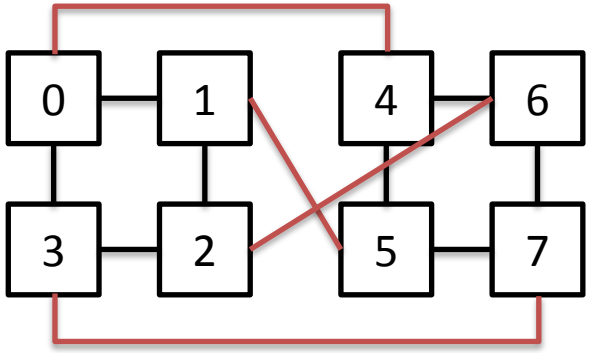
Topology of the 8 Socket Intel Nehalem

Measuring Distance Information

- **Our distance tests:**
  - 8 Threads per socket
  - bandwidth was measured
  - created load on the other sockets
  - scaled to compare to SLIT

Socket	0	1	2	3	4	5	6	7
0	10	12	12	12	12	12	12	12
1	12	10	12	12	12	12	12	12
2	12	12	10	12	12	12	12	12
3	12	12	12	10	12	12	12	12
4	12	12	12	12	10	12	12	12
5	12	12	12	12	12	10	12	12
6	12	12	12	12	12	12	10	12
7	12	12	12	12	12	12	12	10

SLIT for an 8 Socket Intel Nehalem



Topology of the 8 Socket Intel Nehalem

Socket	0	1	2	3	4	5	6	7
0	10	15	18	16	19	26	26	26
1	15	10	15	17	25	19	25	25
2	17	15	10	15	25	25	19	25
3	15	17	15	10	26	25	25	19
4	19	25	25	26	10	15	15	17
5	25	19	25	26	15	10	17	15
6	25	25	19	25	15	17	10	15
7	26	26	25	19	17	15	15	10

self-created distance map on
8 Socket Intel Nehalem

Distance Matrix

Use a Matrix A in the OpenMP runtime for distance information, where A_{ij} describes the distance between Thread i and j.

Proposed ways to initialize a distance matrix:

- **Use the SLIT as base.**
- **Measurements by the OpenMP runtime.**
- **Take a specified file.**
- **Take a user specified function to measure the distance.**

Only useful together with binding!

Threadbinding

OpenMP 3.1:

OMP_PROC_BIND={true|false}

Proposed extension:

OMP_PROCSET={ list of processor IDs }

	OMP_PROC_BIND=true	OMP_PROC_BIND=false
OMP_PROCSET=0,1	Bind T0 to C0 T1 to C1 T2 to C0 …..	Bind T0 to 0,1 T1 to 0,1 T2 to 0,1
OMP_PROCSET nor set	Take all cores the process may use and bind round robin.	Restrict all threads to the cores, the process may use.

Threadbinding

OpenMP 3.1:

OMP_PROC_BIND={true|false}

Proposed extension:

OMP_PROC_BIND={true|false|compact|scatter}

compact: Minimize the sum of distances between all chosen cores.

scatter: Maximize the sum of distances between all chosen cores.

Querying distance information

For “standard” users:

Use compact or scatter strategy to bind threads.

For “expert” users:

```
int omp_get_distance(int a, int b)
```

Returns the distance between Thread a and b.

Evaluation

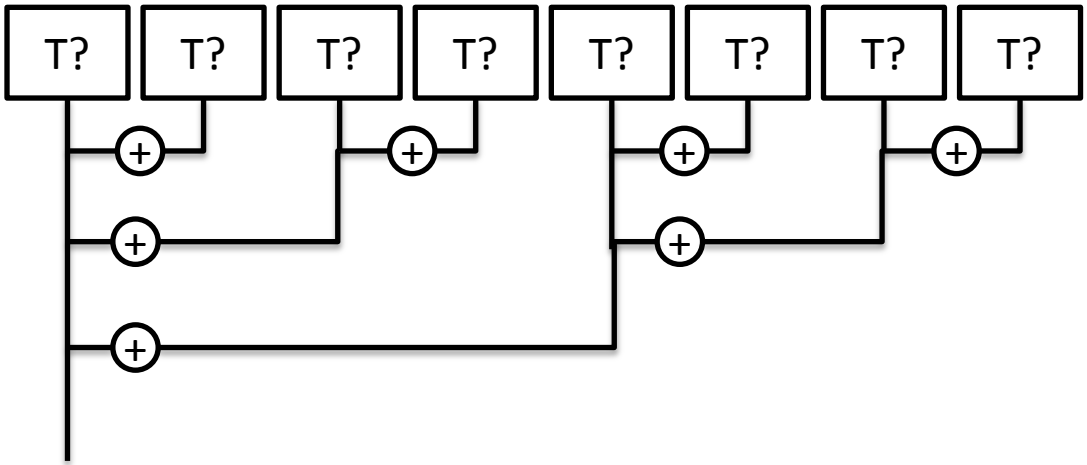
User Code temp1 from Fraunhofer Institute for Laser Technology

- private arrays MA_PRIV used for calculation
- summed up in shared array MA at the end
- MA protected with a critical

Array Reduction

Better approach:

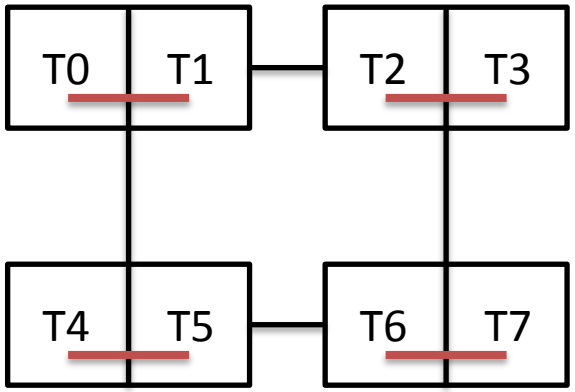
- sum up arrays pairwise



Which threads should sum up the arrays first?

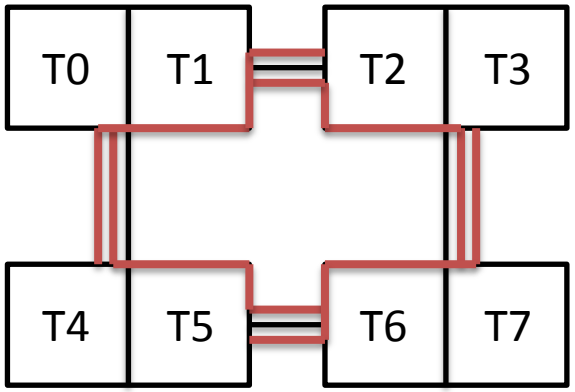
Array Reduction

Solution 1:



Step 1: T0 + T1 ; T2 + T3 ; T4 + T5 ; T6 + T7

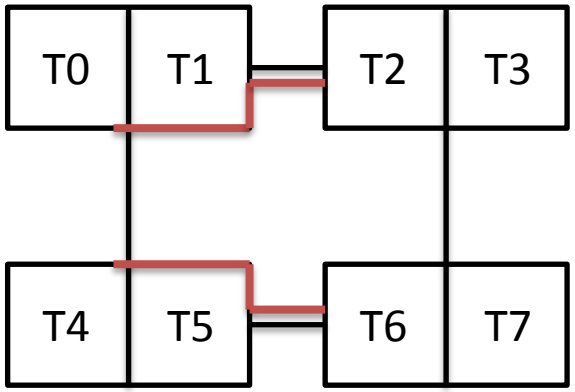
Solution 2:



Step 1: T0 + T6 ; T7 + T1 ; T4 + T2 ; T3 + T5

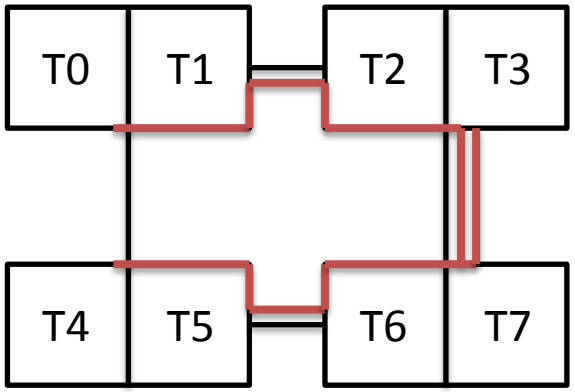
Array Reduction

Solution 1:



Step 1: T0 + T1 ; T2 + T3 ; T4 + T5 ; T6 + T7
Step 2: T0 +T2 ; T4 + T6

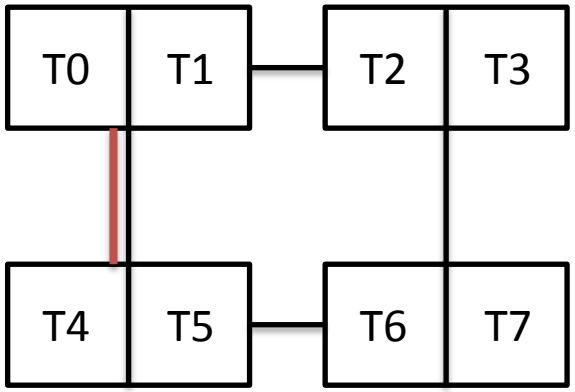
Solution 2:



Step 1: T0 + T6 ; T7 + T1 ; T4 + T2 ; T3 + T5
Step 2: T0 +T7 ; T4 + T3

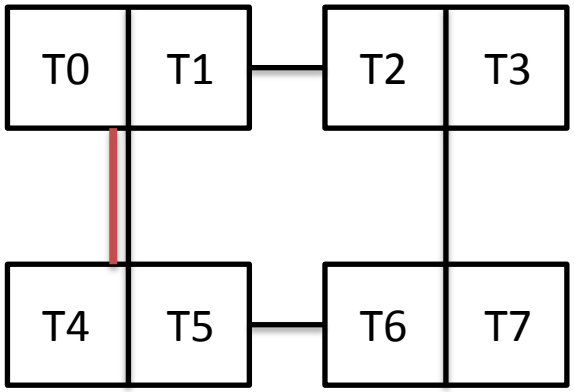
Array Reduction

Solution 1:



Step 1: T0 + T1 ; T2 + T3 ; T4 + T5 ; T6 + T7
Step 2: T0 +T2 ; T4 + T6
Step 3: T0 + T4

Solution 2:

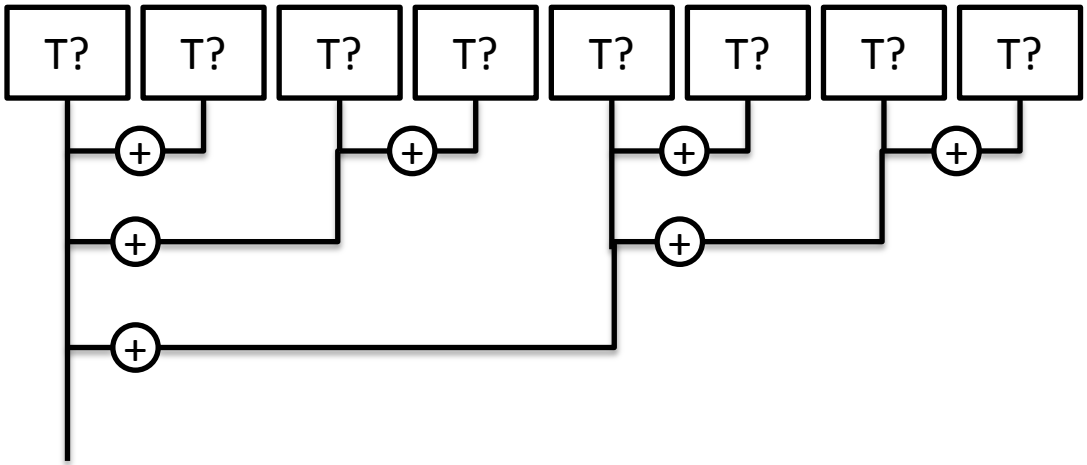


Step 1: T0 + T6 ; T7 + T1 ; T4 + T2 ; T3 + T5
Step 2: T0 +T7 ; T4 + T3
Step 3: T0 + T4

Array Reduction

Better approach:

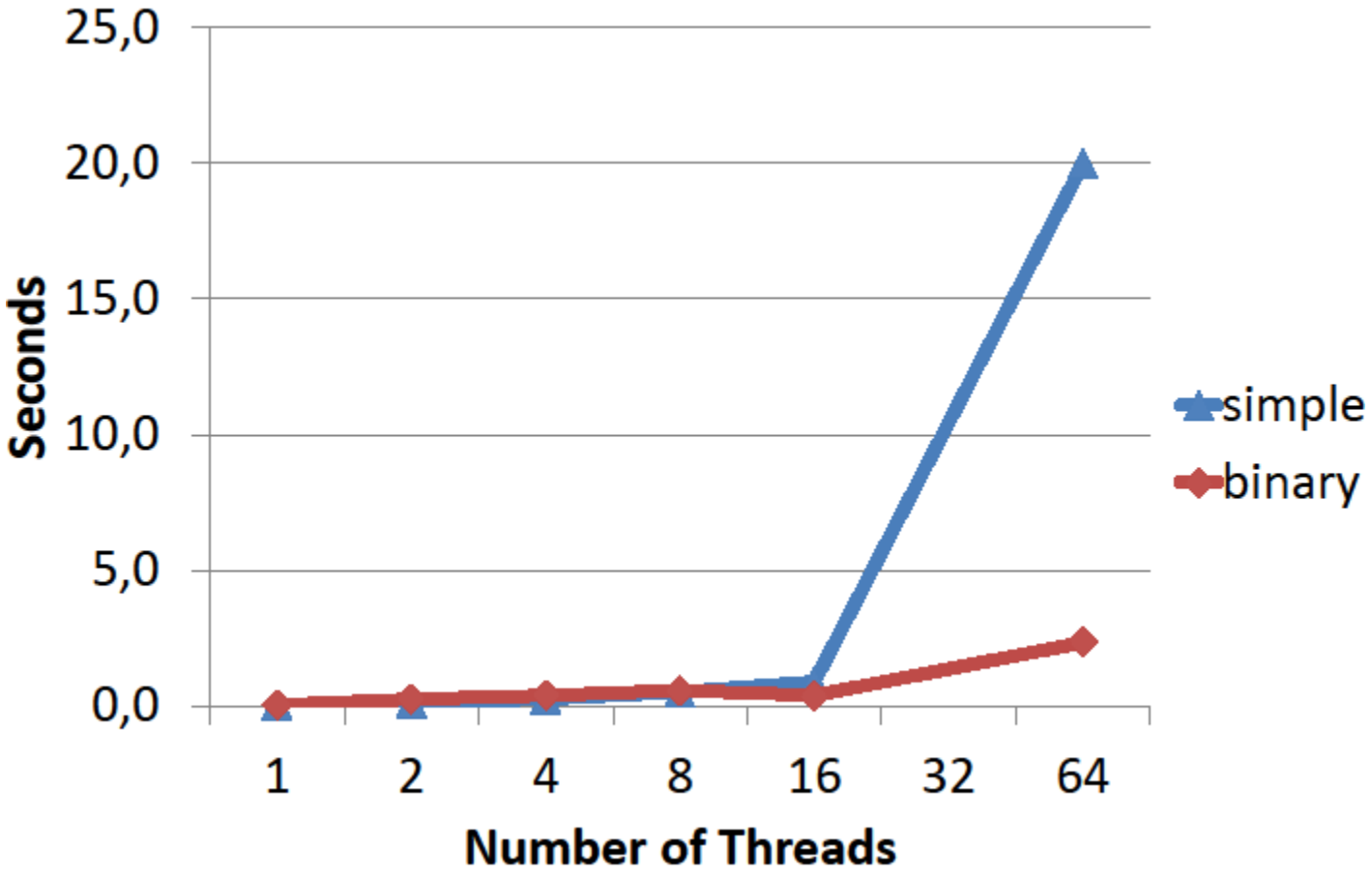
- sum up arrays pairwise



Which threads should sum up the arrays first?

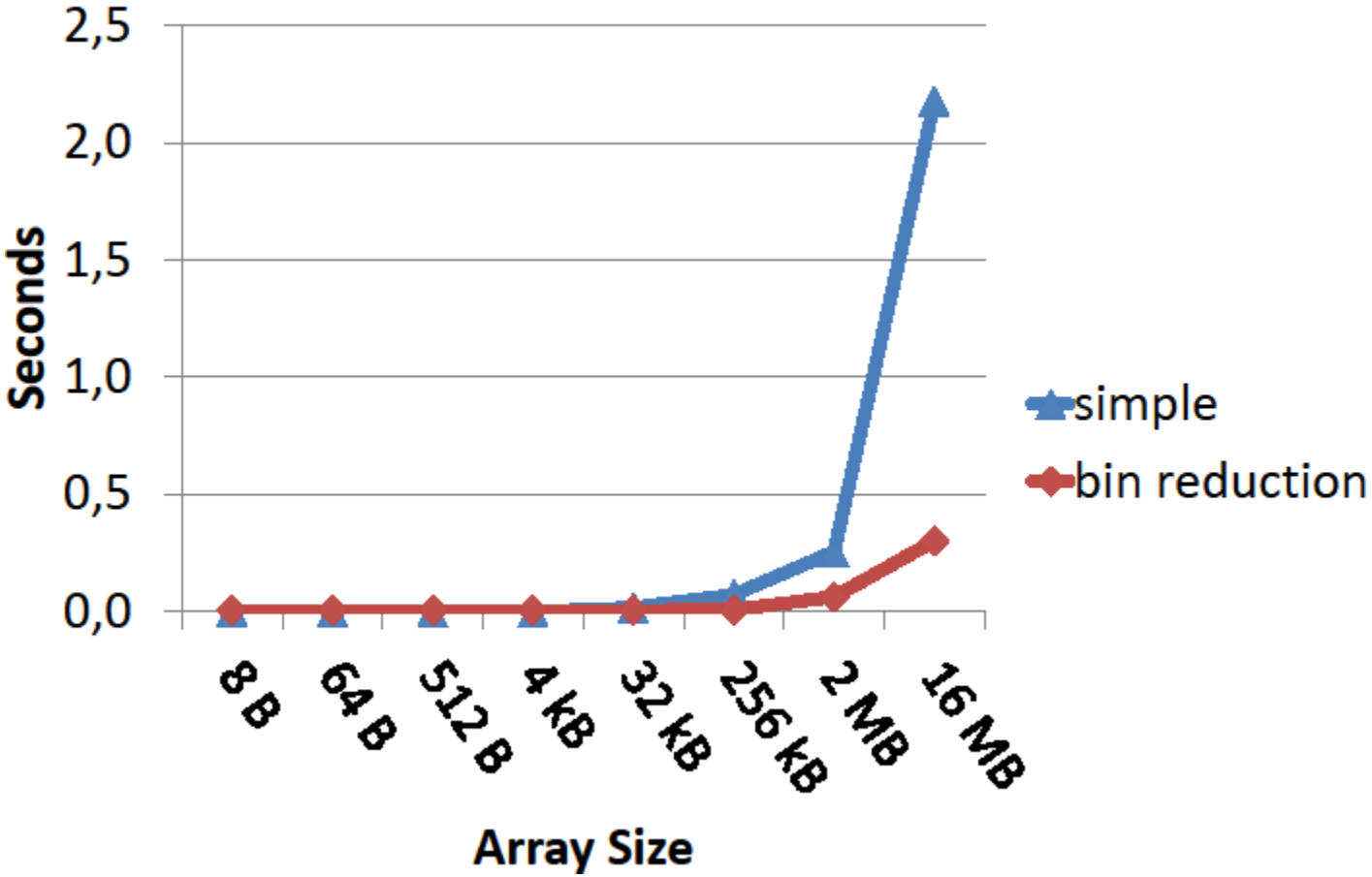
Use the threads with the lowest distance first.

Array Reduction



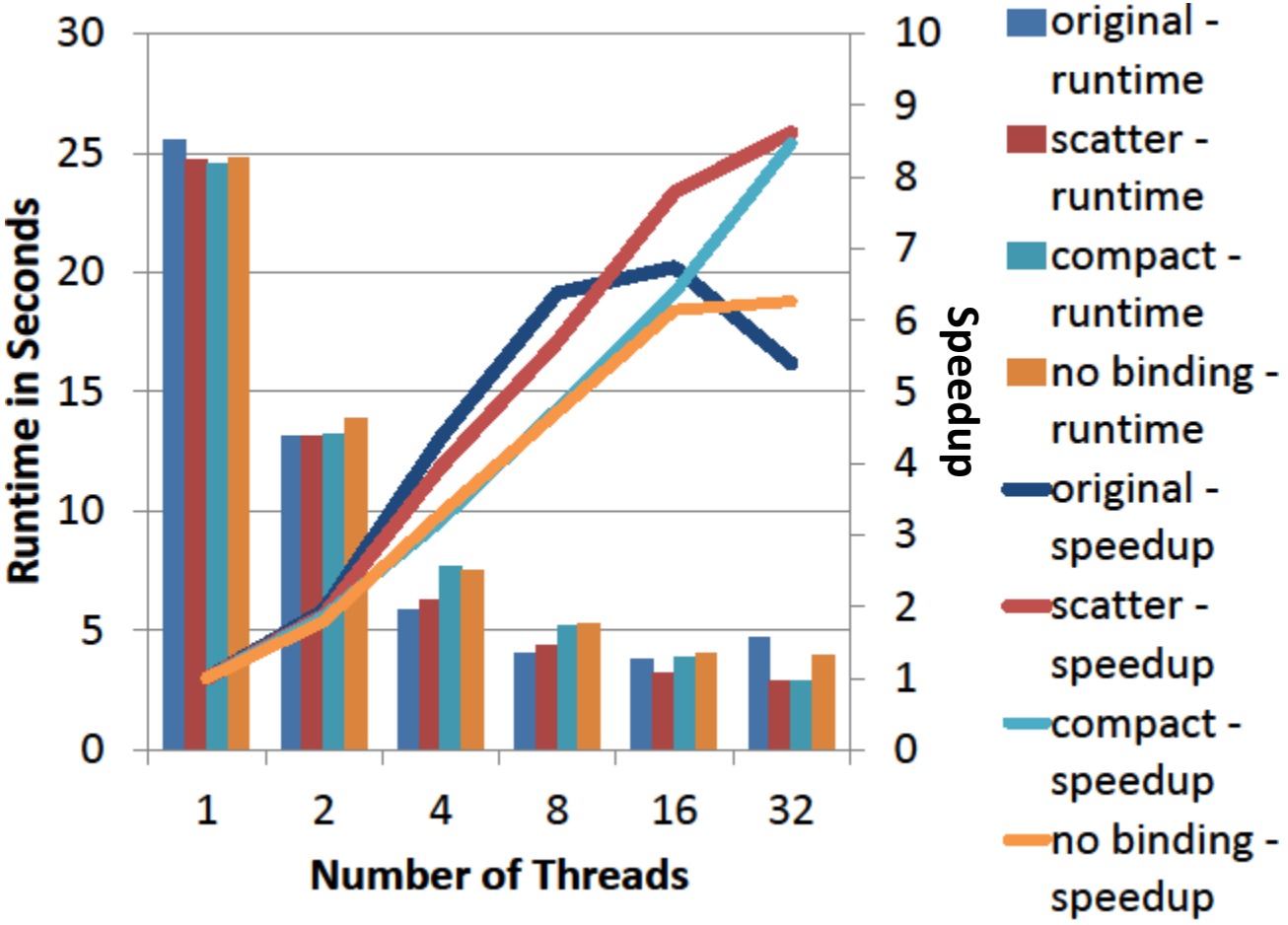
Reduction of 4 MB Arrays on a Altix UV

Array Reduction



Reduction of different sized Arrays on a Altix UV

Temp1



Runtime and Speedup of Temp1 on a 4-Socket Nehalem EX System

Conclusion

- Topologies of modern systems are complicated
- Programmers need to tune for the hardware
- OpenMP should help to do this in an easy way

Proposal:

- OpenMP should provide a distance matrix with distances between threads.
- OMP_PROCSET={ list of processor IDs }
- OMP_PROC_BIND={true|false|compact|scatter}
- int omp_get_distance(int a, int b)