Evaluating the Scalability of Stencil Codes at Scale

*Manish Modani, Rupert Ford
Daresbury Laboratory, UK

Constantinos Evangelinos
IBM Research, USA

(EASC 22/4/2015)

* Presenter

Outline

- Introduction
- Benchmark
- Observation/Issues
- Proposed Mapping
- Energy Measurement
- Conclusions

Communication at Scale

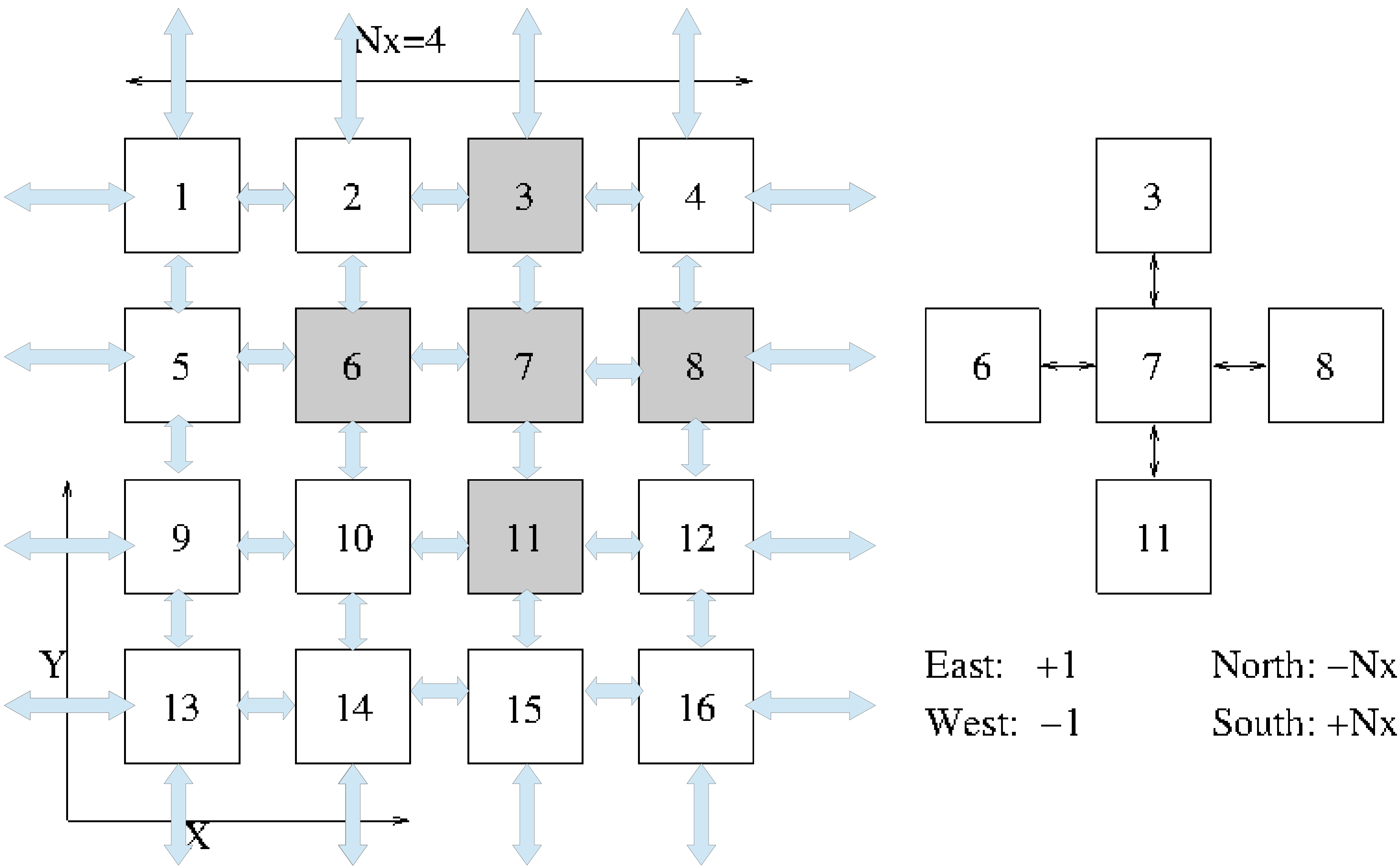
- We have witnessed an ever slower increase and, most recently, reduction in the clock speeds of microprocessor cores.
- This has resulted in higher core counts – the largest systems today have in excess of 100,000 cores and this will increase further for exascale (systems in 1995 had around 512 microprocessors).
- The designing of parallel applications needs to be revisited at all the stages for higher core counts (e.g. Partitioning / Communication /Mapping).
- Communication cost with neighbours is the key for the performance.

Benchmark

- Stencil-based codes are widely used in Scientific Computing and are considered to be good candidates for running at scale.
- 2D stencil kernel has been written to test the assumption that stencil-based codes scale.
- This kernel performs the halo communication that a stencil code would require but does no computation.
- The halo communication is repeated large number of times.

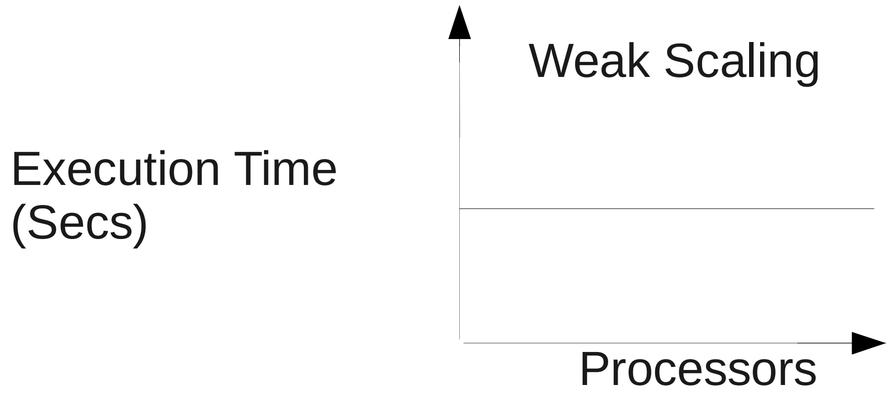
Communication

A (N,E,S,W) communication pattern is used which is representative of a 2 dimensional partitioning of a 5-point stencil on regular grid.



Weak Scaling

- Considered weak scaling to have better idea about the code's scalability.
- Each task communicates with the same number of neighbour's and communicates the same amount of data, irrespective of the number of tasks being used.
- As the computation per task remains the same, the compute to communicate ratio remains the same.



MPI IMPLEMENTATION

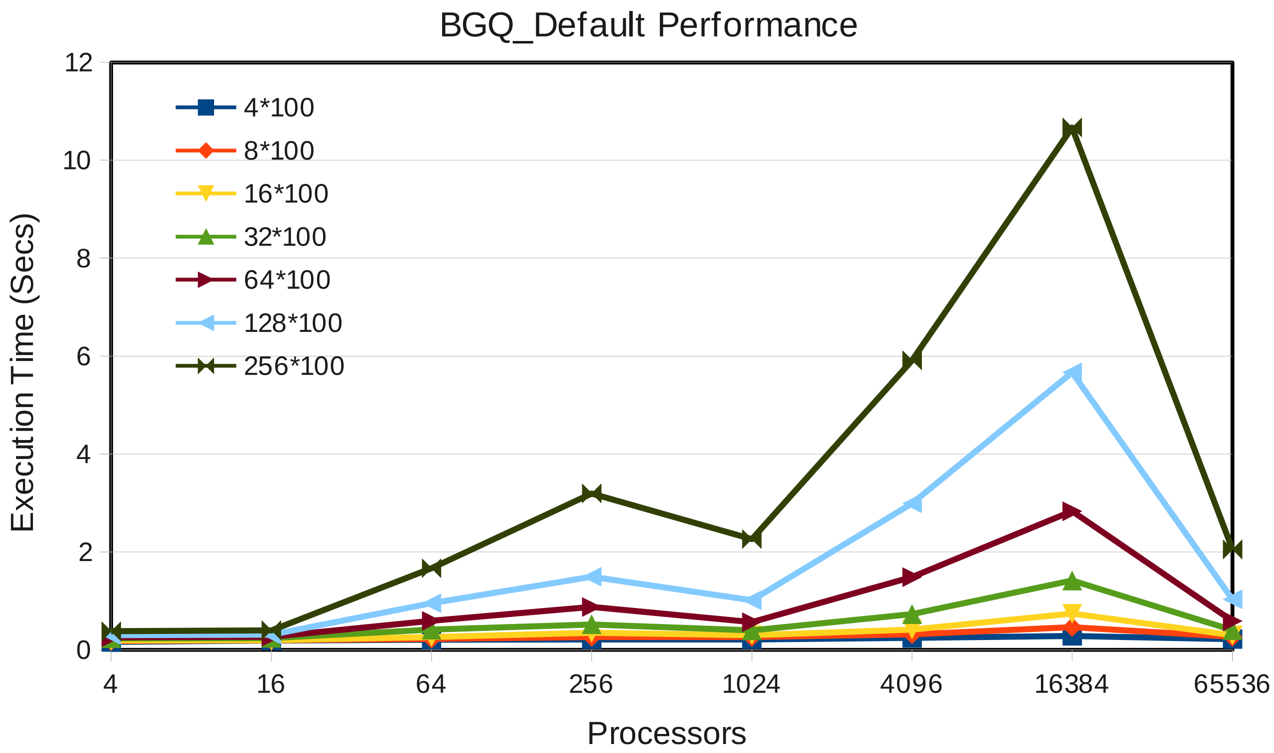
- Communication using MPI.
- MPI Dead-Lock:
  - Even & Odd ranks communication.
- This kernel run using a range of problem sizes on a Blue Gene Q up to 65,000 cores and on TITAN & ARCHER up to 16,000 cores.

Message Size

The halo sizes range from 3,200 bytes per halo (100 levels *4 columns * 8) to 204,800 bytes per halo (100 levels * 256 columns * 8) bytes.

The particular pattern and sizes were chosen as they cover what is used by a large number of Atmosphere models in Climate and Weather Forecasting however the results are relevant to other disciplines.

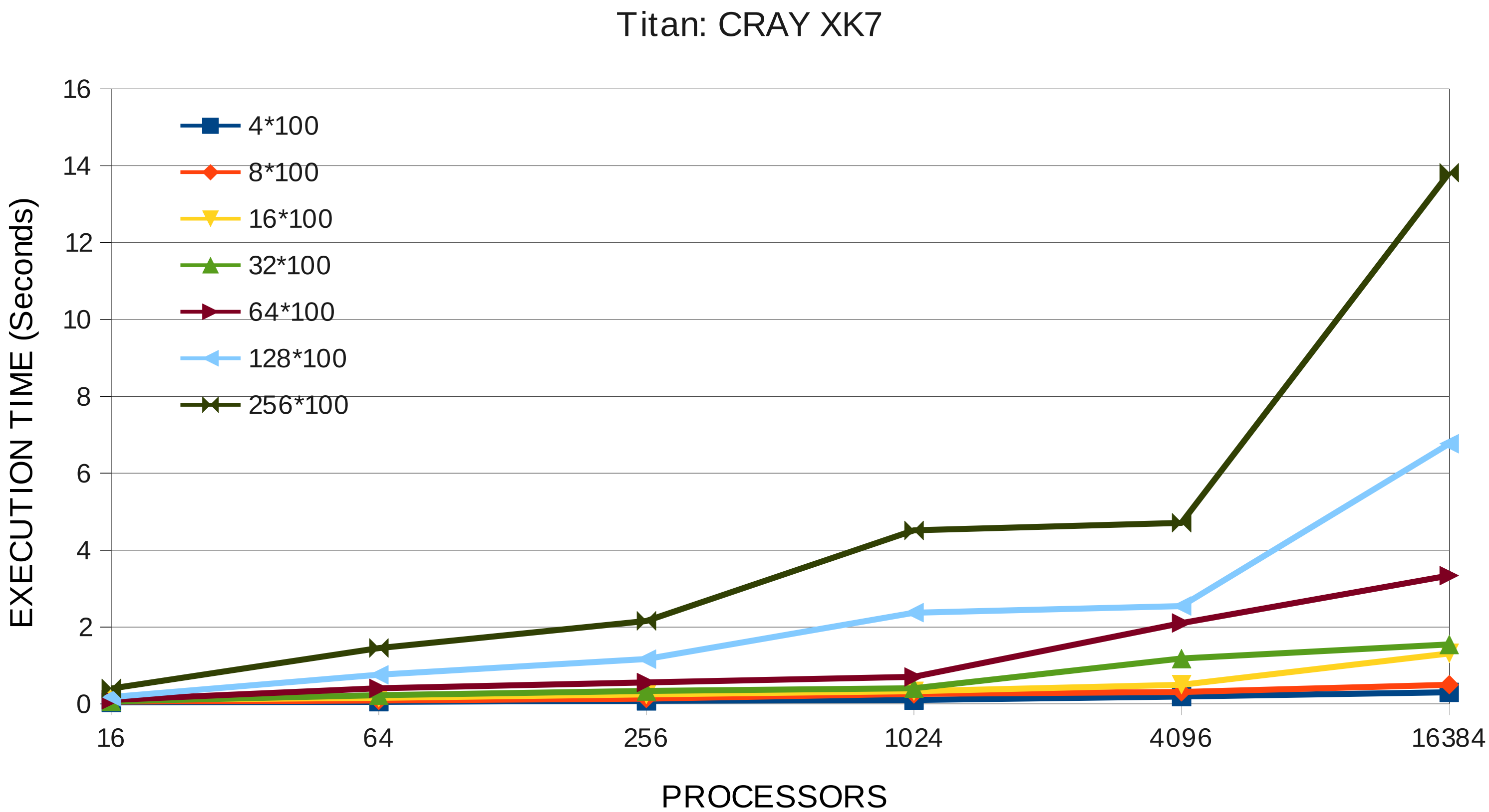
Observations: Blue Joule



Blue Joule - Blue Gene/Q: Hartree Centre UK
Blue Joule consists of 6 racks,
Each rack containing 1,024 nodes
Each node has a 16-core, 64 bit
A2 Power PC, 1.60 GHz processor.
Nodes are interconnected by 5D TORUS

Not Scaling

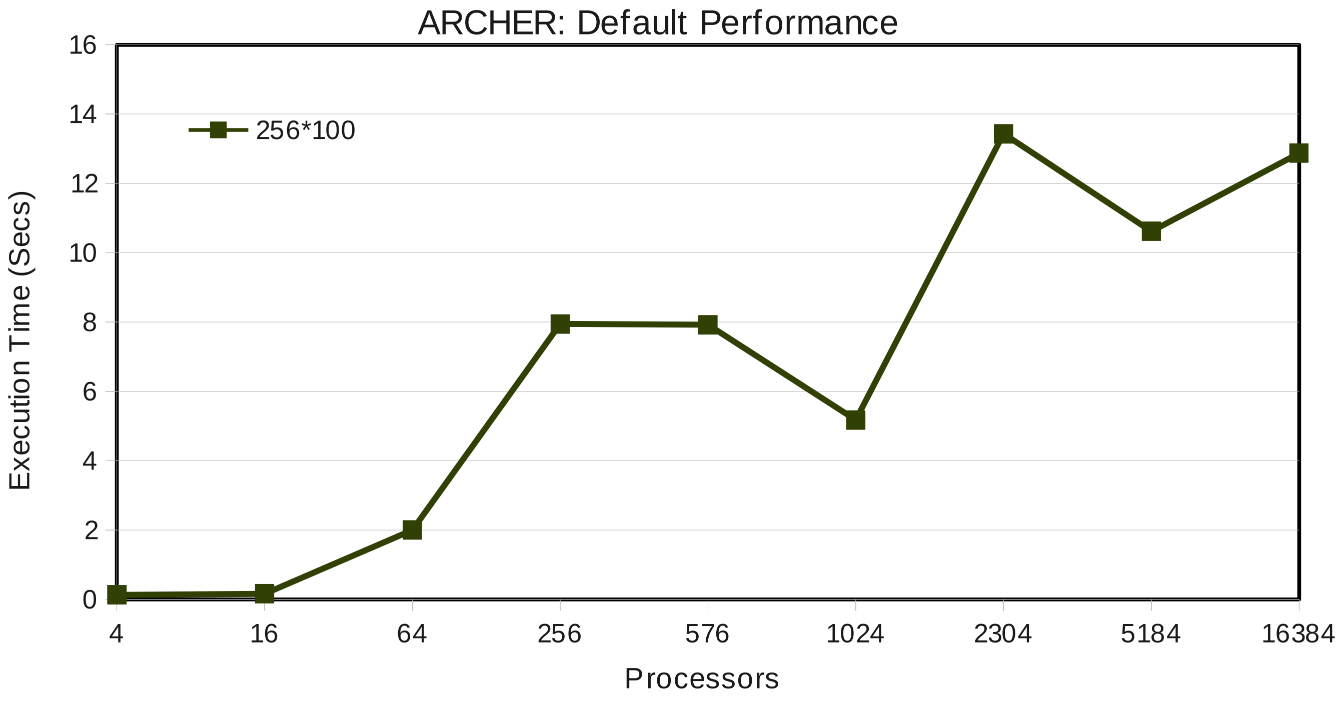
Observation: Titan



Titan: Oak Ridge National Laboratory, USA.
CRAY XK7
18,688 AMD Opteron Cores
16-core CPUs
Gemini Interconnect

Not Scaling

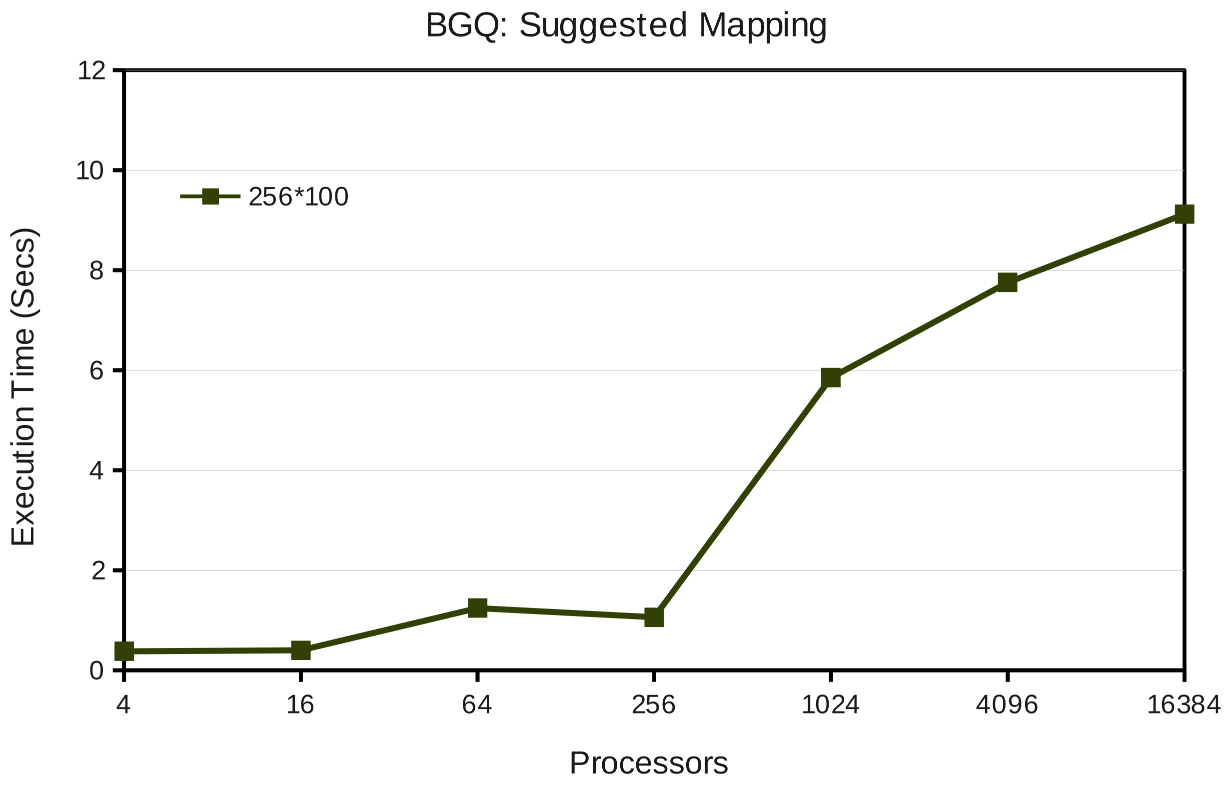
Observation ARCHER



ARCHER UK:
Cray XC30-Architecture,
2.7 GHz Ivy Bridge,
24 cores per node,
Aries Interconnect-Dragonfly topology
Nodes are connected to each Aries router;
188 nodes are grouped into a cabinet;
and two cabinets make up a group

Not Scaling

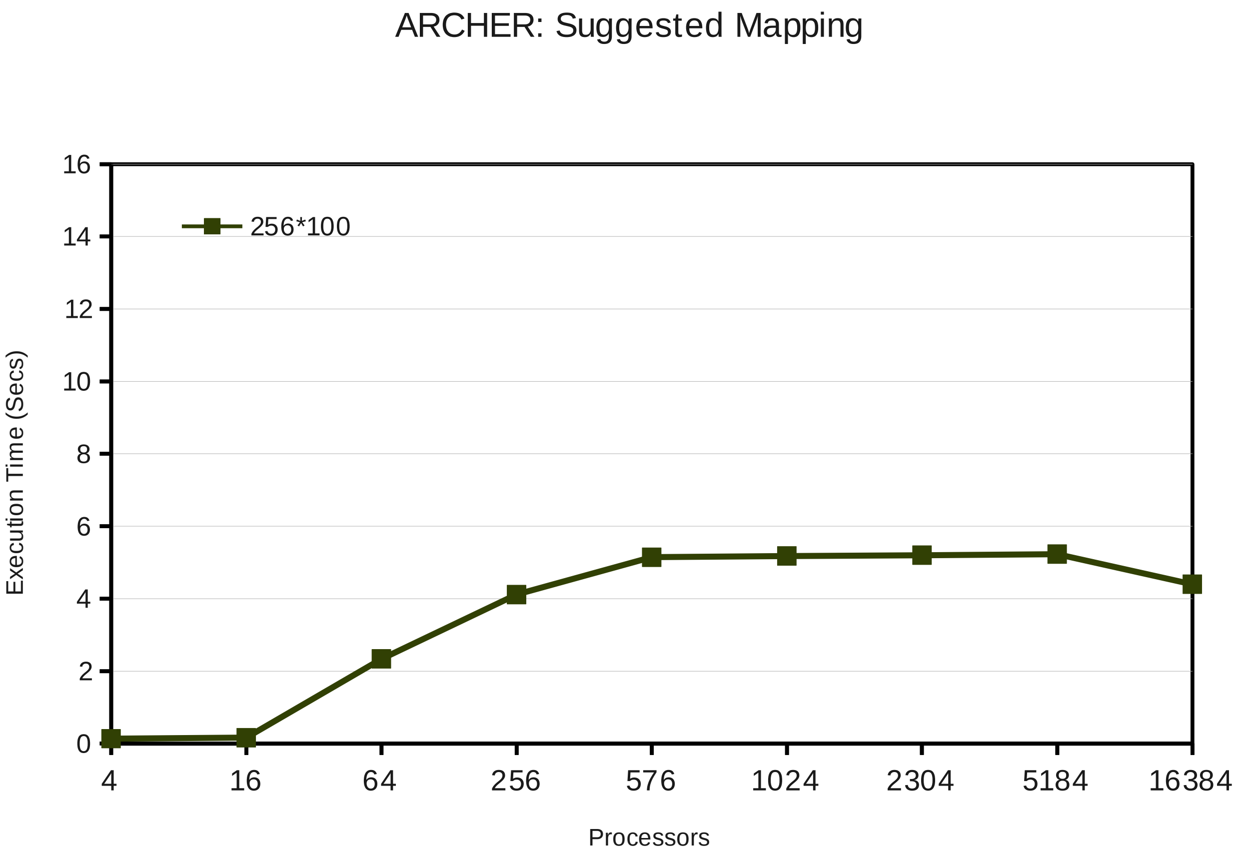
BGQ Tool: Cartesian Topology



Cartesian-Topology Utility to map MPI Ranks on TORUS Network

Requires information about the shape of the block.

ARCHER: Perftools suggested mapping



Cray performance tool
Perftools Module: Using Pat_build and Pat_report a mpi rank file generated.
Rerun the code with new rank file using the variable MPICH_RANK_REORDER_METHOD=3

Analysis

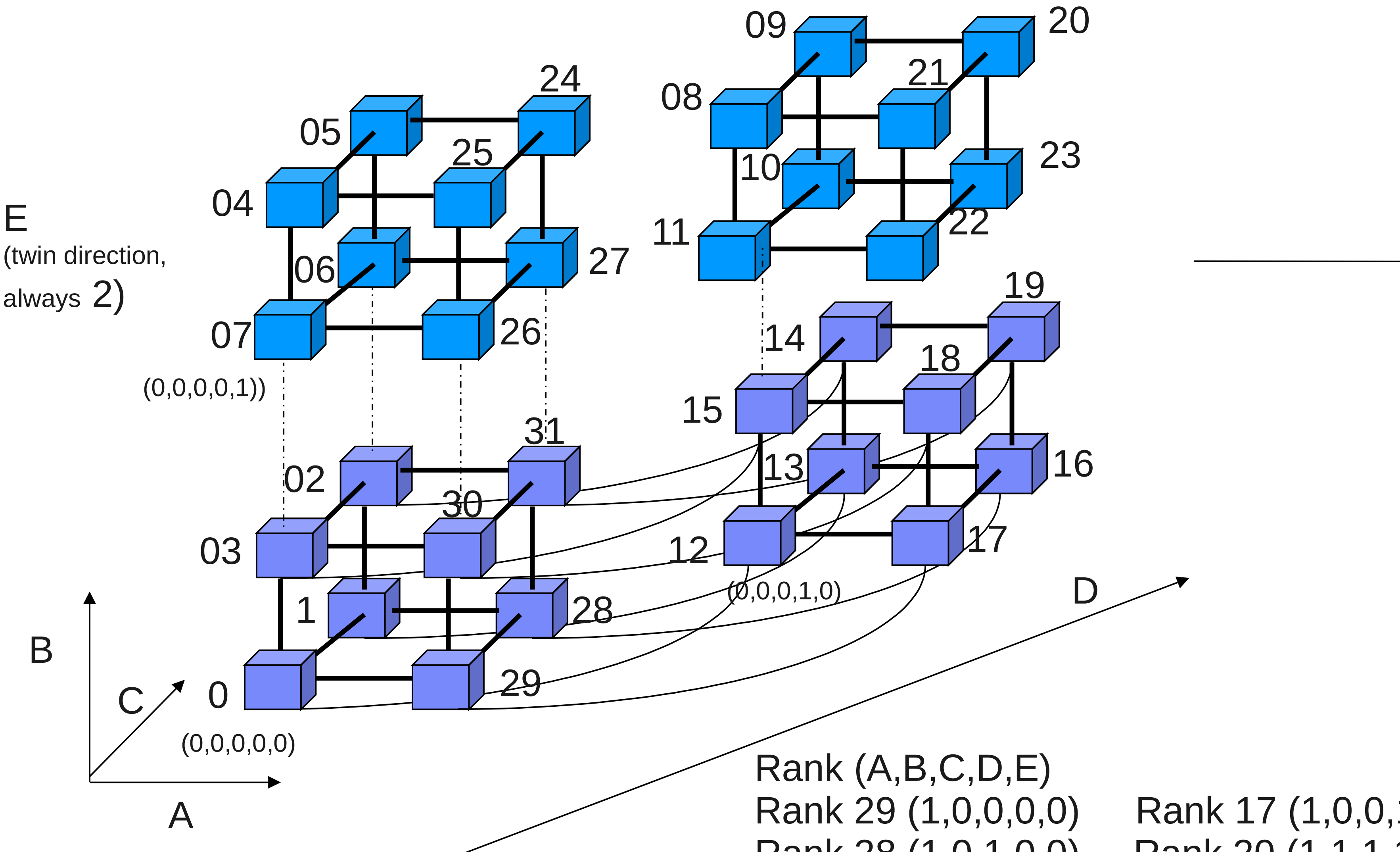
The performance of the code was analysed for 1 dimensional.

It was observed that the communication is X directions scales well.

In Y direction it does not, where most of the messages travels outside the node.

For Y direction, message needed to travel many hops.

5DTORUS : BGQ
Node Board (32 Compute Nodes): 2x2x2x2x2



Rank (A,B,C,D,E)
Rank 29 (1,0,0,0,0)
Rank 28 (1,0,1,0,0)
Rank 31 (1,1,1,0,0)
Rank 17 (1,0,0,1,0)
Rank 20 (1,1,1,1,1)

Task-to-Topology Mapping: Intra-Node Communication

- Default Layout: 16x16: 256 MPI task

240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255
224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239
208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223
192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207
176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191
160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127
96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111
80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95
64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Default: 32 communication outside node

Intra-Node Communication

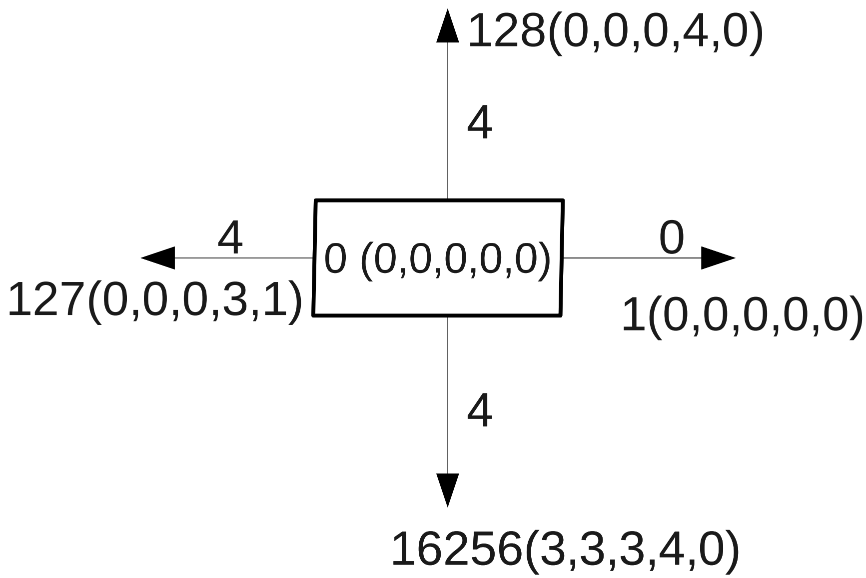
- Minimum communication outside a node

204	205	206	207	220	221	222	223	236	237	238	239	252	253	254	255
200	201	202	203	216	217	218	219	232	233	234	235	248	249	250	251
196	197	198	199	212	213	214	215	228	229	230	231	244	245	246	247
192	193	194	195	208	209	210	211	224	225	226	227	240	241	242	243
140	141	142	143	156	157	158	159	172	173	174	175	188	189	190	191
136	137	138	139	152	153	154	155	168	169	170	171	184	185	186	187
132	133	134	135	148	149	150	151	164	165	166	167	180	181	182	183
128	129	130	131	144	145	146	147	160	161	162	163	176	177	178	179
76	77	78	79	92	93	94	95	108	109	110	111	124	125	126	127
72	73	74	75	88	89	90	91	104	105	106	107	120	121	122	123
68	69	70	71	84	85	86	87	100	101	102	103	116	117	118	119
64	65	66	67	80	81	82	83	96	97	98	99	112	113	114	115
12	13	14	15	28	29	30	31	44	45	46	47	60	61	62	63
8	9	10	11	24	25	26	27	40	41	42	43	56	57	58	59
4	5	6	7	20	21	22	23	36	37	38	39	52	53	54	55
0	1	2	3	16	17	18	19	32	33	34	35	48	49	50	51

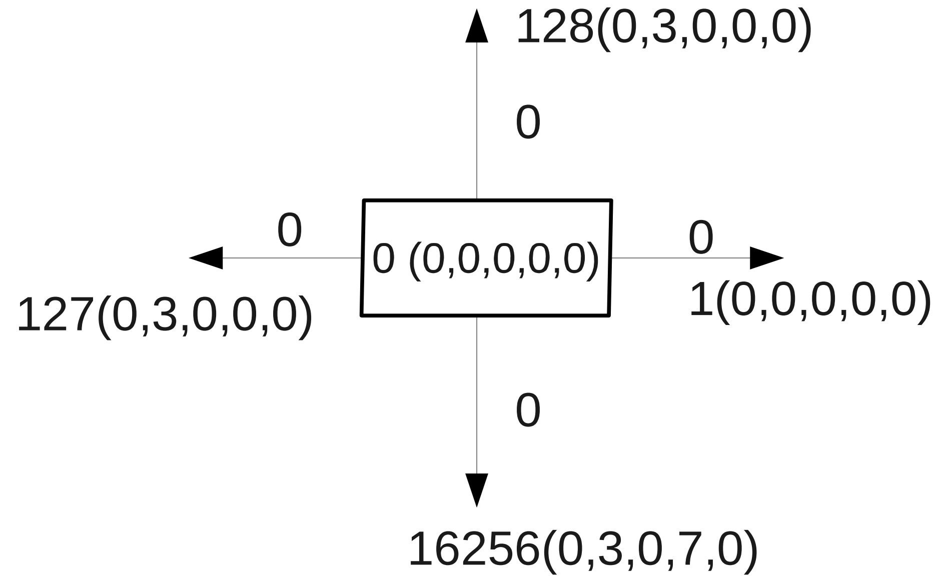
Proposed: 16 communication outside node

Task-to-Topology Mapping

- MPI Ranks re-ordered to travel < = 1 hop
  for 4x4x4x8x2 (128x128; 16384 mpi tasks)



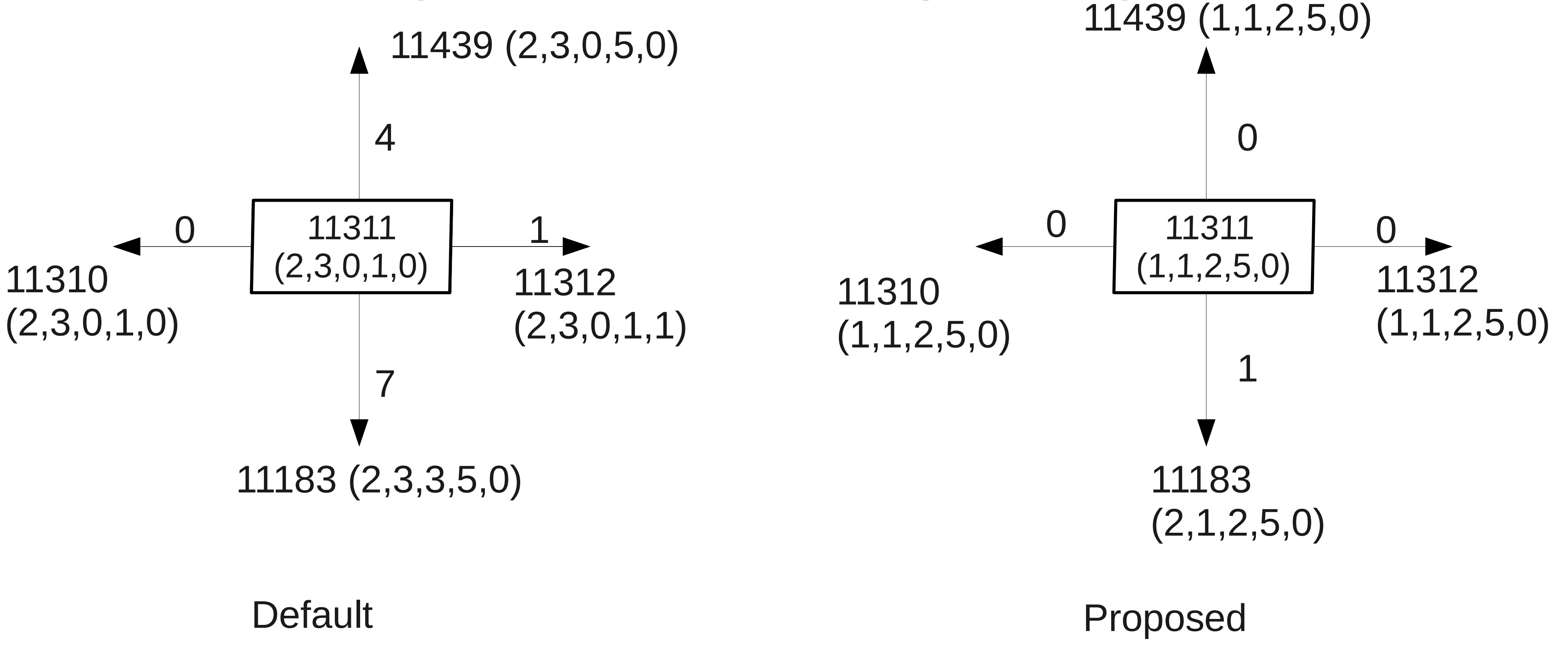
Default



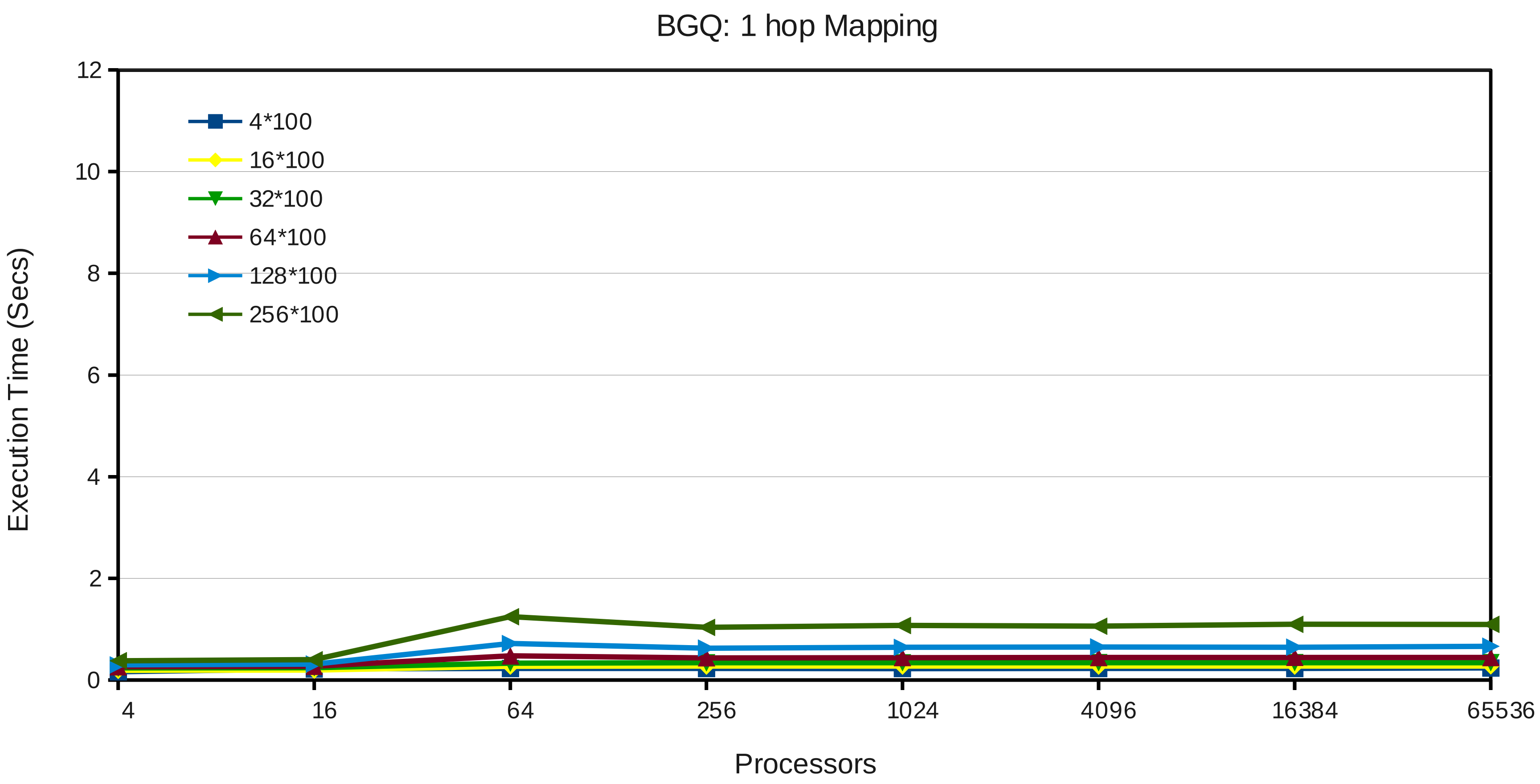
Proposed

Task-to-Topology Mapping

- MPI Ranks re-ordered to travels < = 1 hop e.g. for 4x4x4x8x2 (128x128; 16384 mpi tasks)



Results

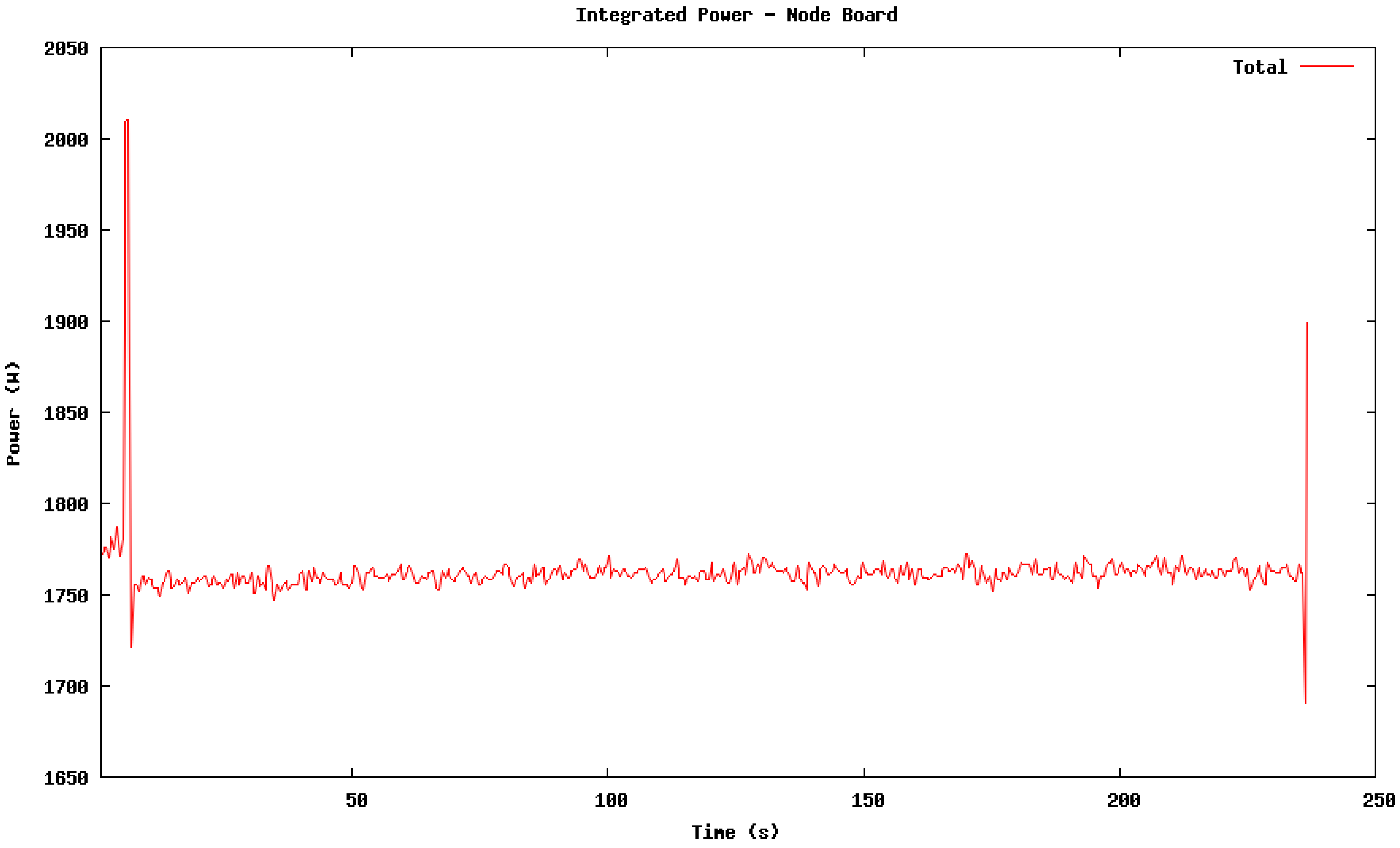


BGQ: 1 hop Mapping
Execution Time (Secs)
0
2
4
6
8
10
12
4*100
16*100
32*100
64*100
128*100
256*100
4
16
64
256
1024
4096
16384
65536
Processors

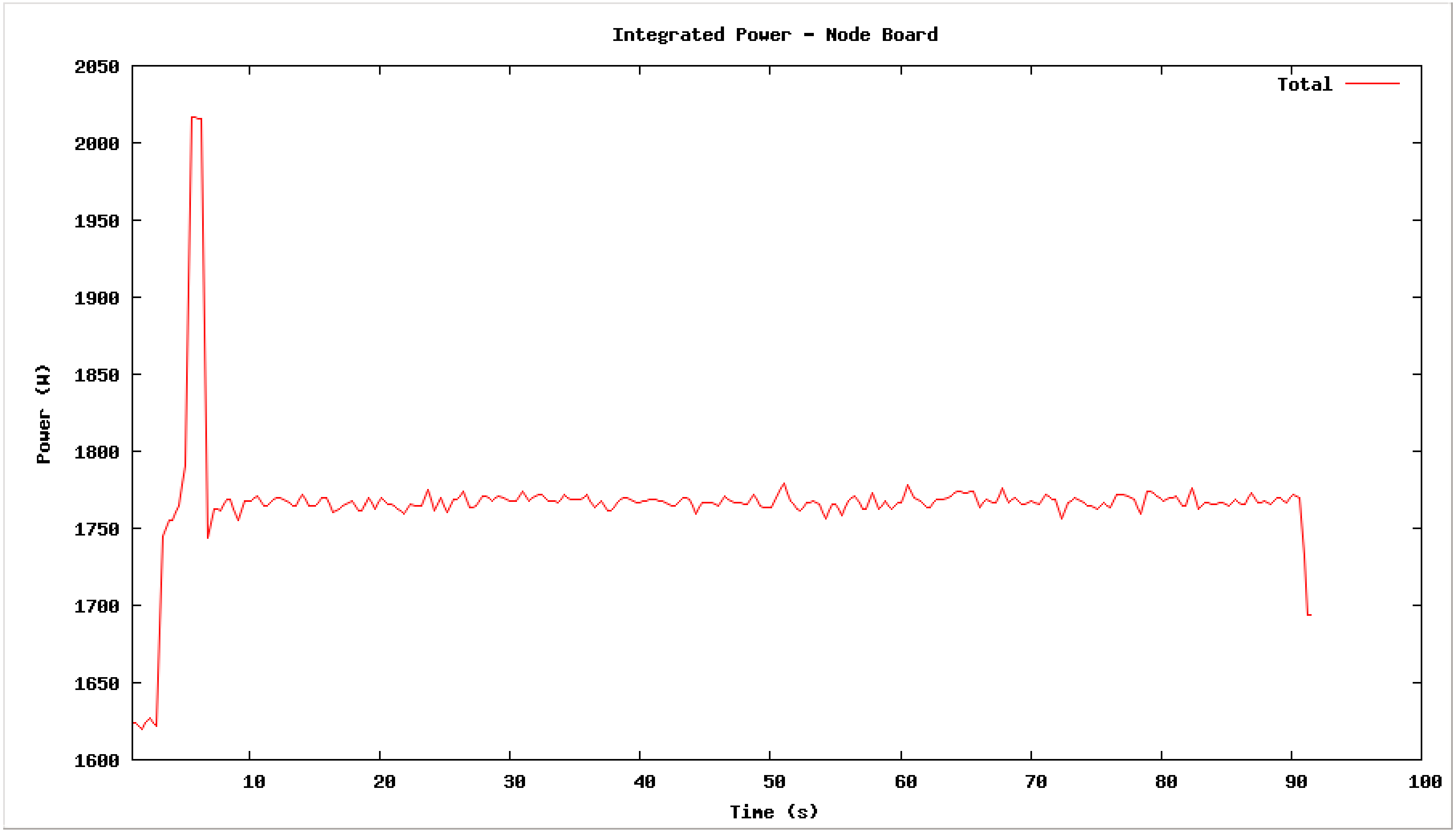
Energy Measurement: BGQ

- EMONSimple is a simple energy monitoring library for Blue Gene/Q. It provides a trace of power consumption versus time for an executable.
- Each BG/Q Node-Board (collection of 32 BG/Q nodes) contains an FPGA that records instantaneous power consumption.
- The sampling frequency of the FPGA is ~ 0.3s and the sampled information can be read by a program via the BG/Q's EMON API.
- This energy information is output at the end of the program giving a trace of the programs energy consumption.
- Easy to use: Link EMON library to binary.

BGQ-Energy: Default



BGQ-Energy:Mapping



Conclusions

- Scalability for weak scaling for 2D-communication to neighbours at scale were observed at scale.
- The default and system tools suggested rank placement does not benefit.
- All the messages shares the same link hence the contention for link bandwidth is the issue.
- 1hop task-to-toppology mapping scheme devised.
- Energy measurement shows upto 60 % less energy consumption from the devised mapping.

Future Direction

Analysis/New Mapping for 3D stencil code.

Acknowledgments

We acknowledge the support provided by Hartree, ORNL & ARCHER Team to use their HPC systems.

Thanking you