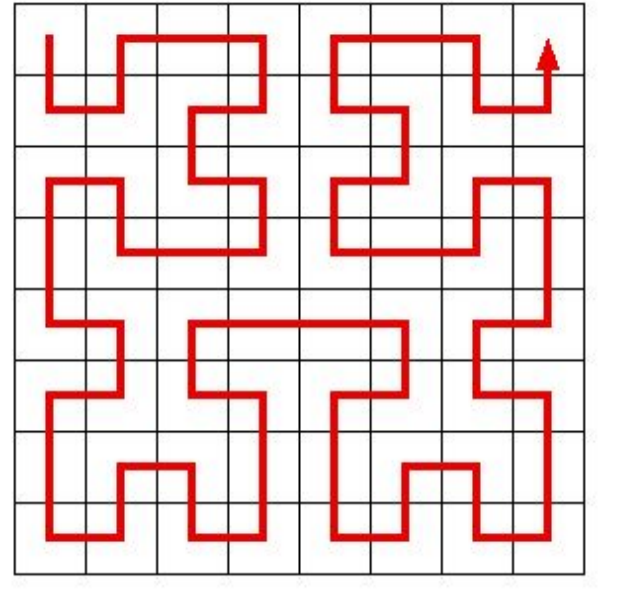




Partitioning the Cubed-Sphere for BlueGene/L



John M. Dennis^{†,‡}, Henry M. Tufo^{†,‡}, Richard D. Loft[‡]

[†]Department of Computer Science,
University of Colorado at Boulder

[‡]Computational Science Section,
National Center for Atmospheric Research

Summary

We present a performance model that predicts the execution time of the primitive equation dynamical core of the NCAR High Order Multi-scale Modeling Environment (HOMME) on $\mathcal{O}(1K)$ of an IBM P690 cluster. The performance model is based on the HOMME communication pattern and fundamental machine characteristics. We predict the execution time and floating-point execution rate for the explicit version of HOMME on $\mathcal{O}(100K)$ BG/L processors for a 9.5 km mesh and compare it to a 10 km mesh atmospheric general circulation model on the Earth Simulator (AFES). The execution rate of HOMME, which should reach 30 - 40 Tflops on 55296 BG/L processors, compares favorably to the 26.58 Tflops achieved by the AFES model on 5120 processors of the Earth Simulator.

Computational Grid

The cubed-sphere computational domain is displayed in Figure 1, where each cube face contains an array of $N_e \times N_e$ quadrilateral spectral elements. The total number of spectral elements is $K = 6 \times N_e \times N_e$. Communication is determined by neighboring spectral elements that share a boundary or corner point.

We examine two cubed-sphere resolutions: a low resolution with $K=1536$ ($N_e = 16$) total spectral elements and a high resolution with $K=55296$ ($N_e = 96$) total spectral elements. We use the $K=1536$ resolution to validate the performance model on an IBM P690 cluster, while the $K=55296$ resolution is comparable to the 10 km mesh of the AFES model on the Earth Simulator [3]. The characteristics of both cubed-sphere meshes are located in Table 1.

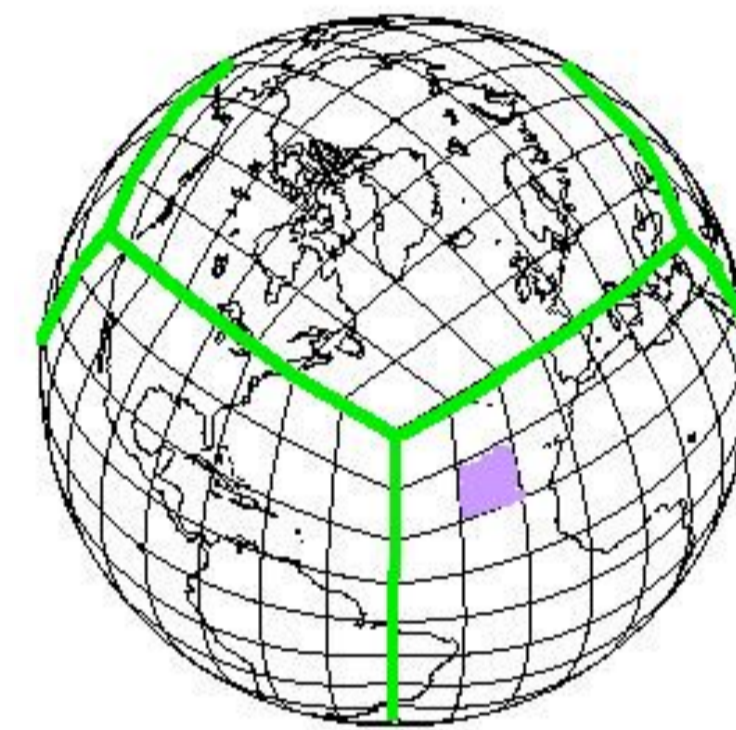


Figure 1: The cubed-sphere with continental outlines for $N_e = 8$. The purple patch is a spectral element, while the green lines indicate edges of cube faces.

K	Ne	vertical levels	spec. elem. order	p-grid points	v-grid points	explicit dt (sec)	equ. grid ΔX (km)
1536	16	16	6	55296	75264	50.0	89.3
55296	96	96	10	5.5 M	6.7 M	5.0	9.5

Table 1: HOMME computational grid sizes.

Space-Filling Curves

An inverse space-filling partitioning algorithm (ISP) is used to partition the computational mesh across processors. We concentrate on partitioning a single face of the cube of size $M \times M$. Our ISP is based on three different curves: a Hilbert curve [2] for $M = 2^n$, a meandering Peano (m-Peano) [2] for $M = 3^m$, and a combined Hilbert and m-Peano curve [1] for $M = 2^n 3^m$. A combined Hilbert and m-Peano curve for $M = 6$ is illustrated in Figure 2. The beginning and end of the space-filling curve on each face is aligned with the curves on adjoining faces to construct a single continuous space-filling curve. A flattened cube with a level 1 Hilbert curve mapped onto the cube is displayed in Figure 3, along with a perspective drawing of the same image. The space-filling curve is then subdivided into equal sized segments to achieve the partitioning.

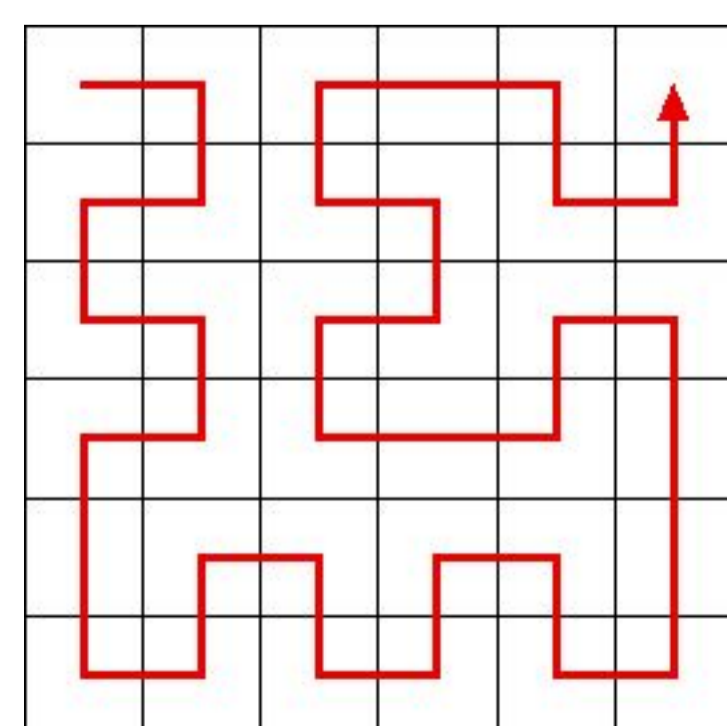


Figure 2: A combined Hilbert and m-Peano curve ($M=6$)

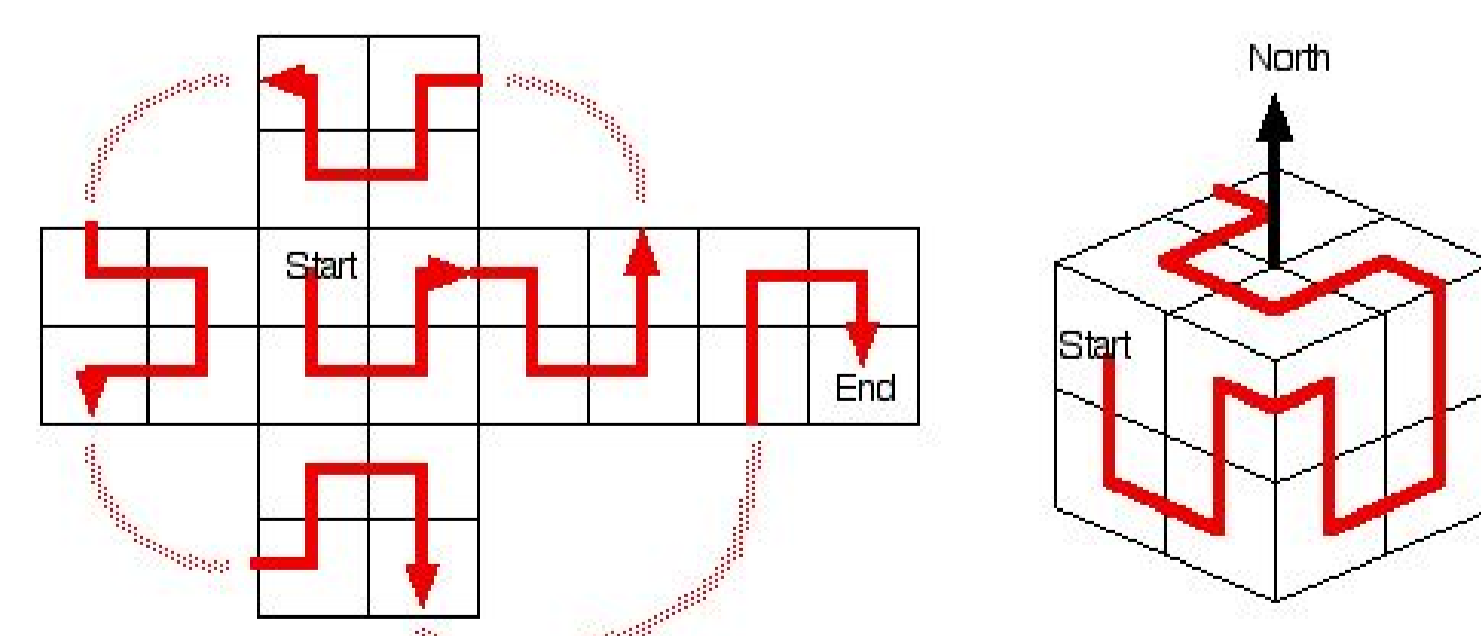


Figure 3: A mapping of a level 1 Hilbert curve onto the flattened cube (left) and perspective drawing (right).

Performance Model

Once the cubed-sphere mesh is partitioned, each partition element is allocated to separate processors. The execution time is the sum of the computation time and the communication time. The computational time for processor k is $D_L \cdot nelemd_k$ where D_L is the time to calculate a single spectral element and $nelemd_k$ is the number of spectral elements on the k^{th} processor. The communication cost is the sum of the time to pass messages between each neighboring processor. The time to execute HOMME on the k^{th} processor is

$$t_k = D_L \cdot nelemd_k + \sum_{l \in Neigh(k)} (ts + s(k, l) \cdot Bw), \quad (1)$$

where $s(k, l)$ is the message volume between the k^{th} and l^{th} processor. Because parallel execution time is $T_{||} = \max_k(t_k)$ and serial execution time is $T_s = D_L \cdot K$ where K is the total number of spectral elements, we can easily calculate the parallel speedup $Speedup = T_s/T_{||}$.

A slightly modified version of (1) is necessary to predict the execution time of HOMME on a P690 cluster. Modifications are required because messages sent between processors on the same SMP node have much lower latency and higher bandwidth than messages sent between processors on different nodes. The execution time for processor k is

$$t_k = D_L nelemd_k + \sum_{l \in Neigh_{off}(k)} (ts_{off} + s(k, l) tBw_{off}) + \sum_{l \in Neigh_{on}(k)} (ts_{on} + s(k, l) Bw_{on}), \quad (2)$$

where the subscripts *off* and *on* indicates off-node and on-node communication costs respectively.

IBM P690 Cluster Results

For the P690 cluster, we concentrate on the $K=1536$ resolution with $L = 16$ vertical levels. Execution times were measured over a range of processor counts, chosen specifically so that an equal number of spectral elements are allocated to each processor. Floating-point rates were calculated based on the values from *hpmcount* for a single processor. The measured speedup and floating-point operation counts are plotted in Figures 3 and 4 along with the predicted values using (2) and the machine characteristics from Table 2. The performance model accurately predicts speedup and execution rate with exception of 768 processors. This discrepancy is because communication costs account for $> 50\%$ of the total time per time step.

Time/ spectral element	Single Processor	off-node	on-node
D_{L16}	880 Mflops	ts_{off}	ts_{on}
$540 \mu s$		Bw	Bw_{on}
		$17.9 \mu s$	$.8 \mu s$
		360 Mbytes/s	1260 Mbytes/s

Table 2: IBM P690 cluster characteristics.

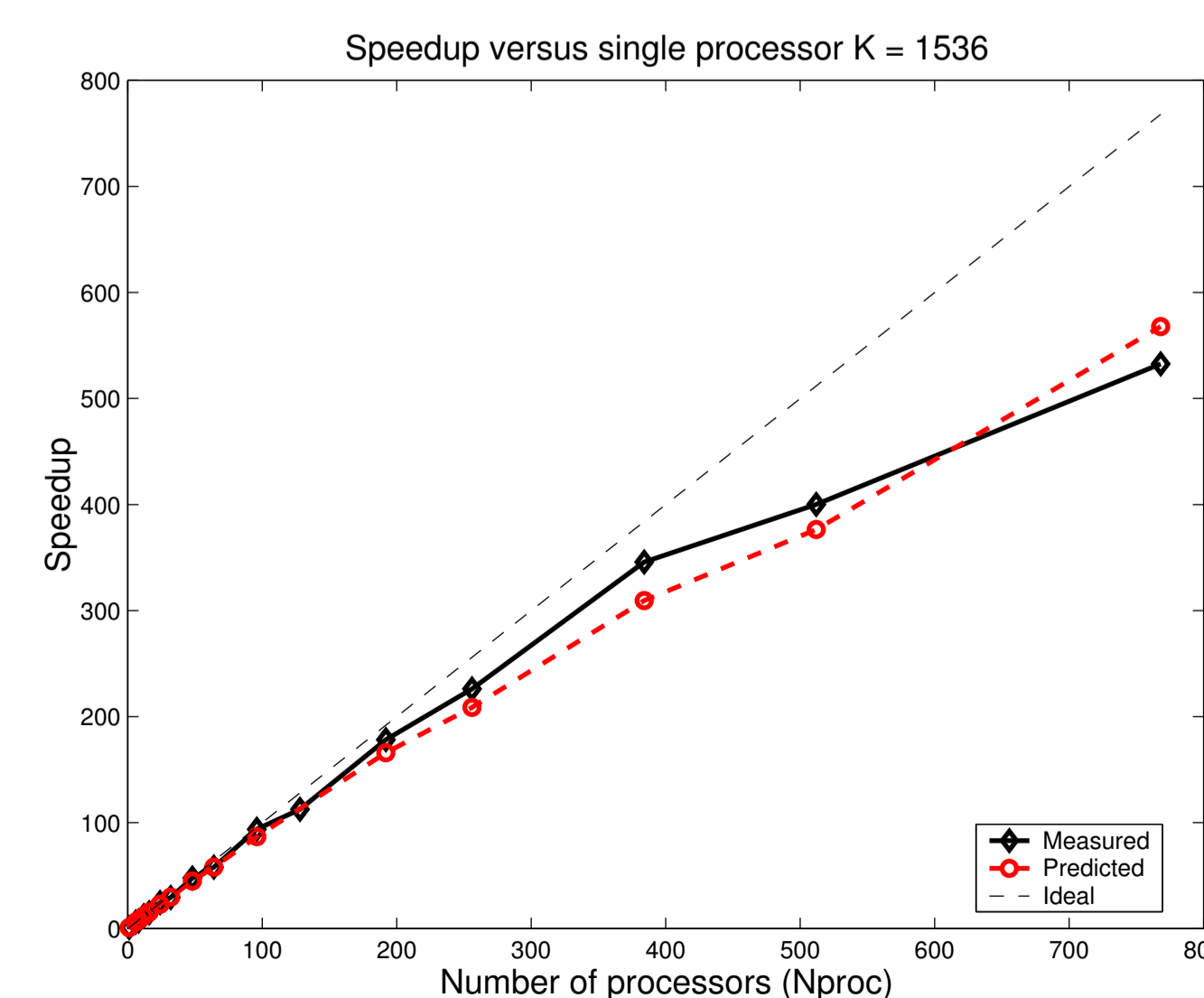


Figure 3: HOMME speedup on P690 cluster.

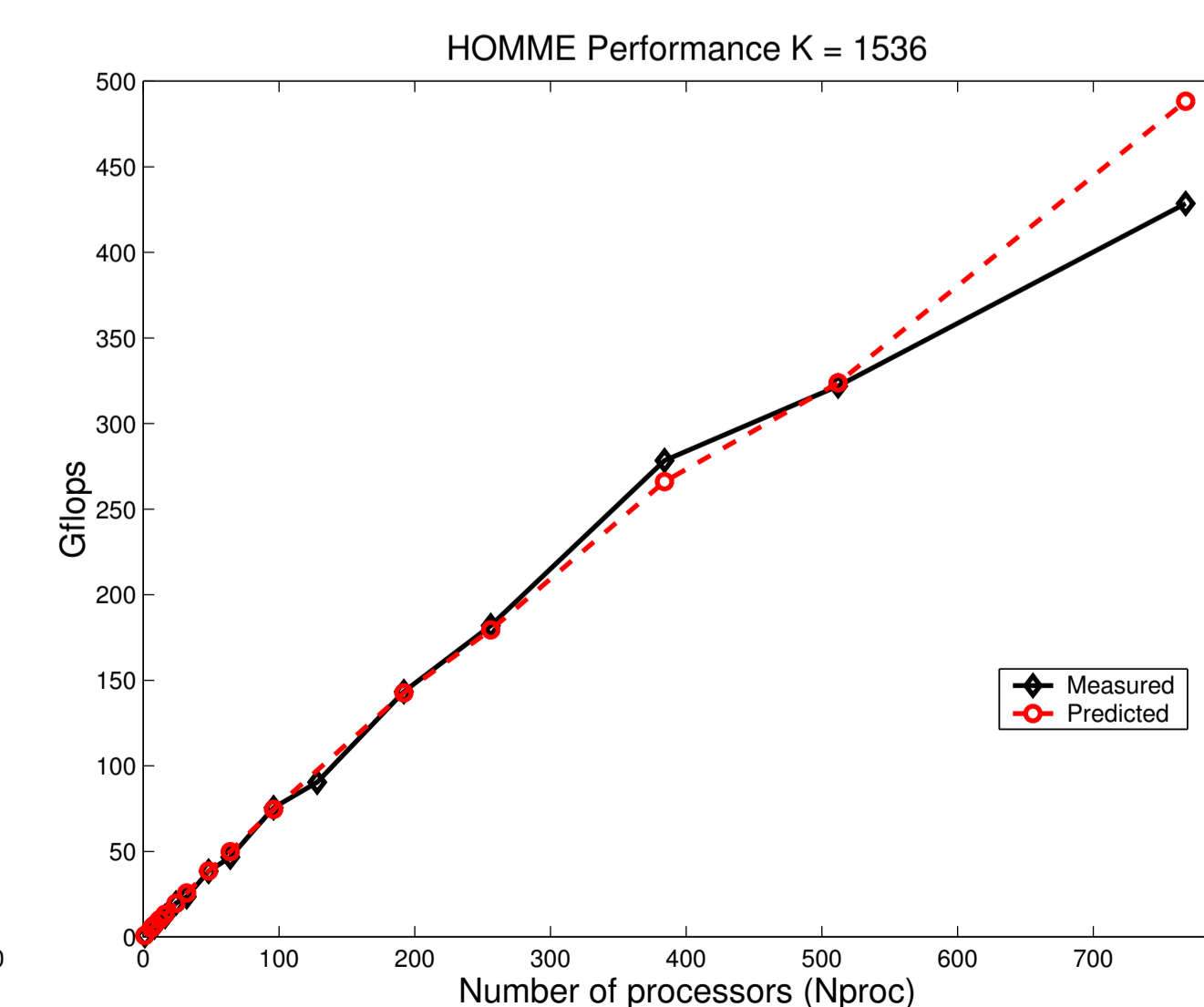


Figure 4: HOMME performance on P690 cluster.

IBM BlueGene/L

We use several different sets of machine characteristics to bound our performance predictions for BG/L. To determine single processor performance, we bound the time to execution a single spectral element (D_{L96}) on the 700 Mhz PowerPC 440 by the execution time on either a Power3 or Power4 processor scaled by clock speed. We therefore estimate a single PowerPC 440 processor to execution HOMME at 550 to 740 Mflops. To calculate the communication cost, we assume the projected BG/L network as an upper bound and a network with $3.0 \mu s$ latency and 85 Mbytes/s per link as a lower bound on network performance. We assume a maximum of 3 messages may content for the same network link because each process has ~ 8 communication neighbors. The four machine configurations are summarized in Table 3.

The predicted speedup versus 1024 processors illustrated in Figure 5, is consistent for all four machine configurations. The nearly ideal speedup is because even at 55296 processors, the communication time is $< 5\%$ of the total time. The total floating-point performance illustrated in Figure 6 also emphasizes the importance of computational costs. The *A* and *D* configurations based on a 740 Mflop processor, are predicted to achieve ~ 40 Tflops while the *B* and *C* configurations should only achieve ~ 30 Tflops.

Configuration	Time/ spectral element D_{L96}	Single processor	Latency ts_{off}	Bandwidth (Bw)	
				per link	total
A	$12973 \mu s$	740 Mflops	$1.5 \mu s$	175 Mbytes/s	1050 Mbytes/s
B	$17467 \mu s$	550 Mflops	$1.5 \mu s$	175 Mbytes/s	1050 Mbytes/s
C	$17467 \mu s$	550 Mflops	$3.0 \mu s$	85 Mbytes/s	510 Mbytes/s
D	$12973 \mu s$	740 Mflops	$3.0 \mu s$	85 Mbytes/s	510 Mbytes/s

Table 3: Several possible BG/L machine characteristics.

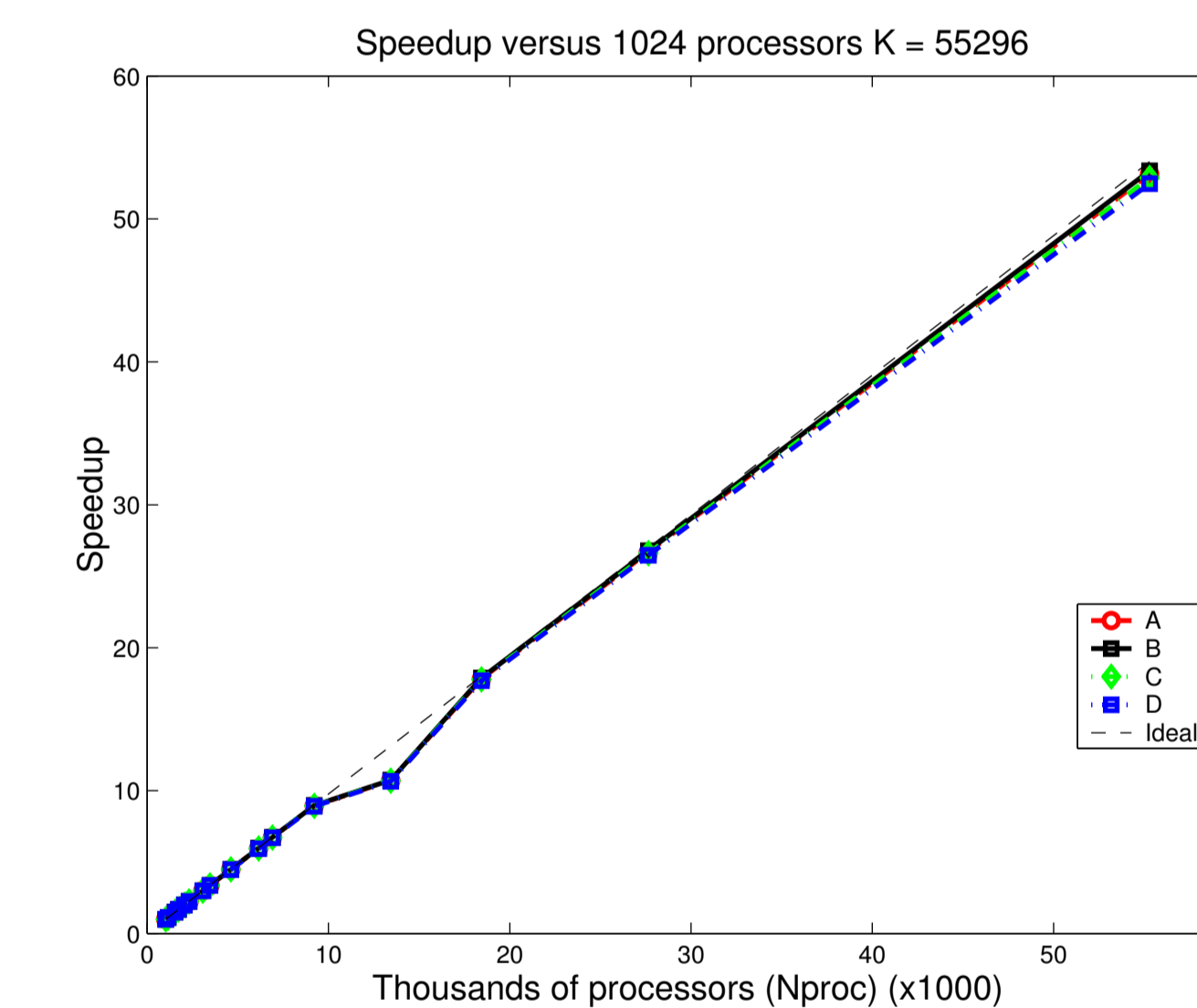


Figure 5: Predicted HOMME speedup on BG/L for four different machine configurations.

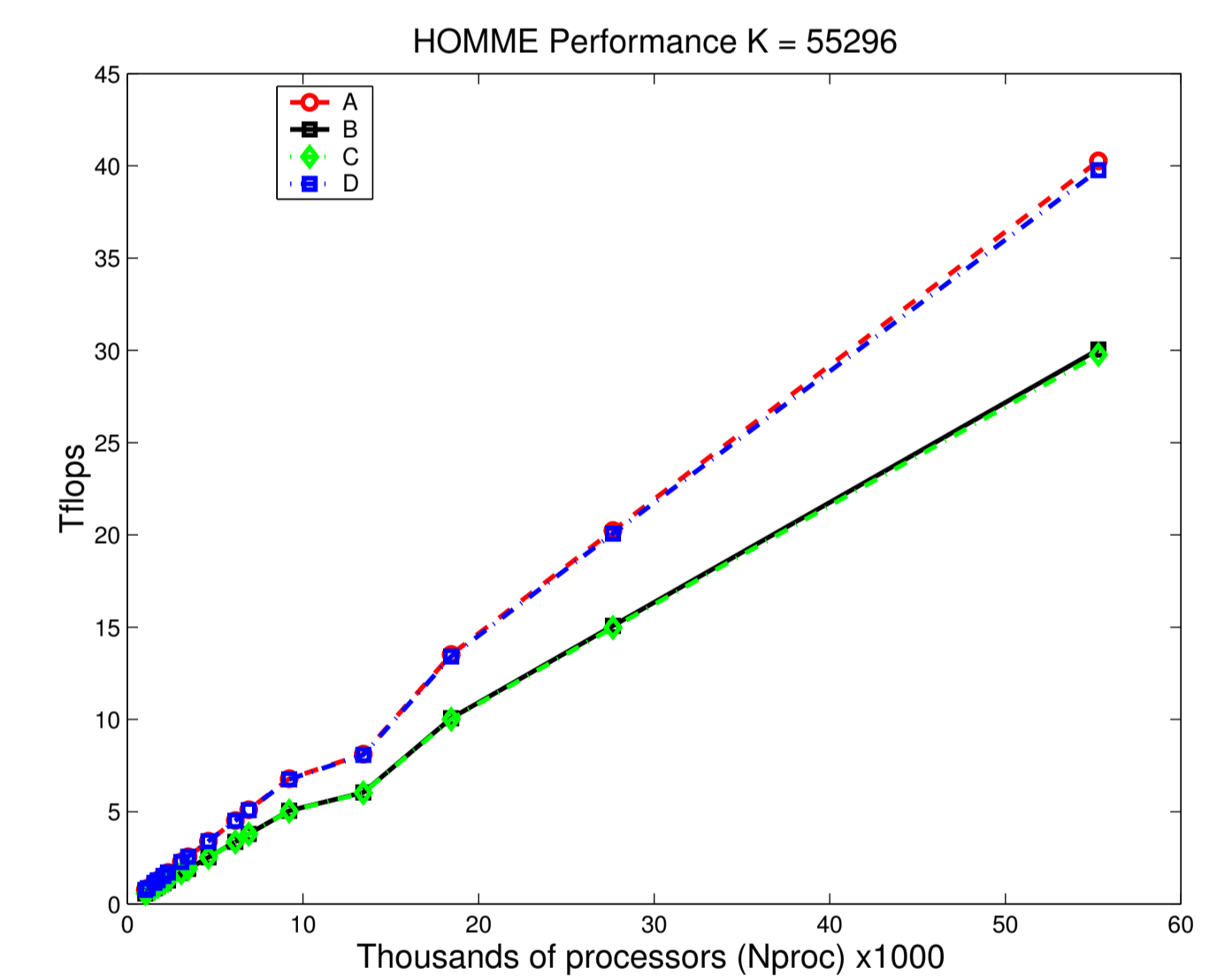


Figure 6: Predicted HOMME performance on BG/L for four different machine configurations.

Conclusions

We demonstrate a performance model that accurately predicts performance of the explicit version of the HOMME model on $\mathcal{O}(1K)$ processors of an IBM P690 cluster when communication costs are $< 50\%$ total time per time step. We predict the performance of a 9.5 km mesh on $\mathcal{O}(100K)$ processors of BG/L. This resolution is dominated by computation cost ($> 95\%$) at 55296 processors, and should achieve 30 to 40 Tflops. This result compares favorably with a similar resolution atmospheric general circulation model that achieved 26.58 Tflops on 5120 processors of the Earth Simulator.

References

- [1] John M. Dennis. Partitioning with space-filling curves on the cubed-sphere. In *Workshop on Massively Parallel Processing at IPDPS, Nice, France, April 2003*
- [2] Hans Sagan. *Space-Filling Curves*, Springer-Verlag, 1994
- [3] S. Shinug and H. Takahara and H. Fuchigami and M. Yamada and Y. Tsuda and W. Ohfuchi and Y. Sasaki and K. Kobayashi and T. Hagiwara and S. Habata and M. Yokokawa and H. Itoh and K. Otsuka, *A 26.58 Tflops Global Atmospheric Simulation with the Spectral Transform Method on the Earth Simulator*, in Proceedings of Supercomputing '02, November 2002