SAND2013-4277C

Application Explorations for Future Interconnects

Photos placed in horizontal position with even amount of white space between photos and header

Photos placed in horizontal position with even amount of white space between photos and header

Richard Barrett, Courtenay Vaughan, Si Hammond, Sandia
Duncan Roweth, Cray, Inc.

Large-Scale Parallel Processing (LSPP) workshop@IPDPS
May 24, 2013
Boston MA, USA

Richard F. Barrett
Scalable Computer Architectures
Sandia National Laboratories, NM
rfbarre@sandia.gov

SAND 2013-TBD.

Sandia National Laboratories

Exceptional service in the national interest

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Photos placed in horizontal position with even amount of white space between photos and header

Photos placed in horizontal position with even amount of white space between photos and header

Sandia National Laboratories

Exceptional service in the national interest

Reducing the Bulk of the Bulk Synchronous Parallel Programming model

Richard Barrett, Courtenay Vaughan, Si Hammond, Sandia
Duncan Roweth, Cray, Inc.

Large-Scale Parallel Processing (LSPP) workshop@IPDPS
May 24, 2013
Boston MA, USA

Richard F. Barrett
Scalable Computer Architectures
Sandia National Laboratories, NM
rfbarre@sandia.gov

SAND 2013-TBD.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Overview

- The problem: Just as apps are scaling to hundreds of thousands of processors, interconnect bandwidth is (proportionally) decreasing.
- Our explorations for addressing.
- Some results.
- Where this is going

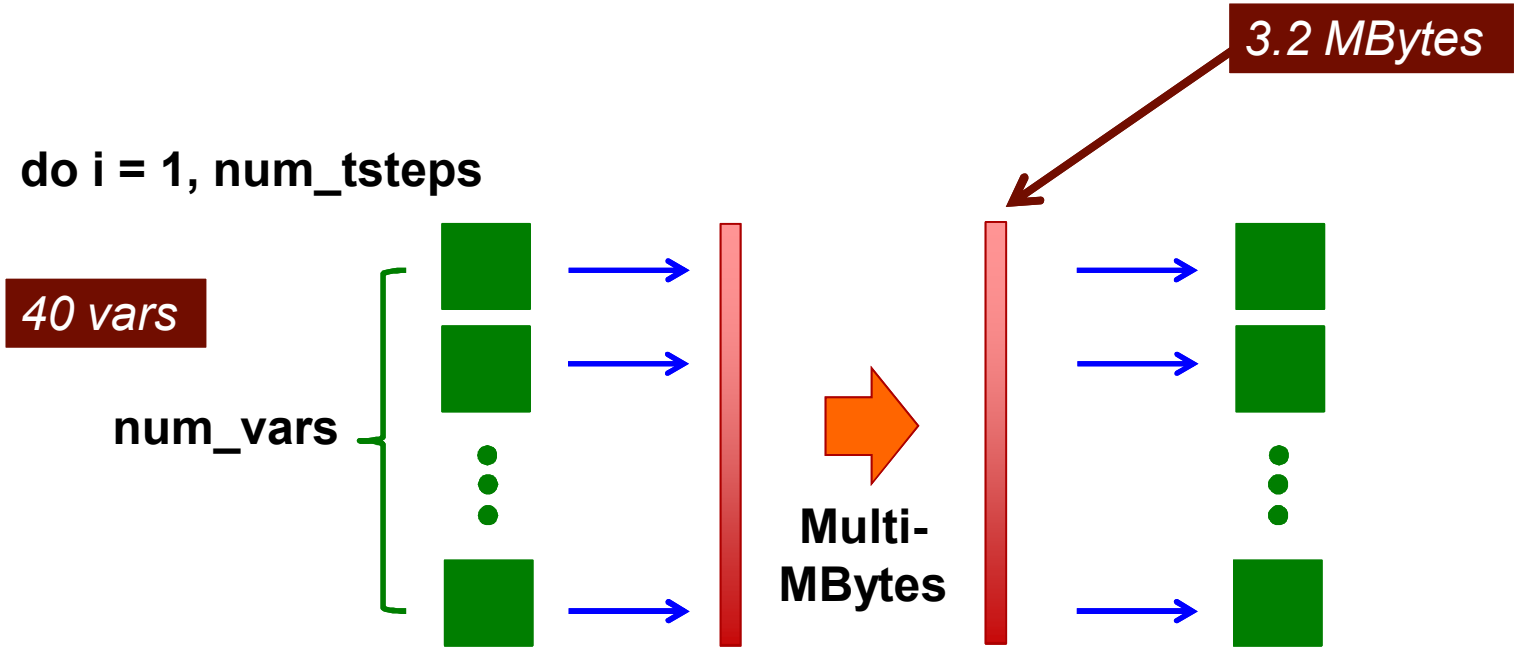
Refining on “the problem”

Sandia National Laboratories

- We’ve been trained to exploit bandwidth by aggregating data into fewer messages.
- Energy limitations are constraining interconnect global bandwidth
- PGAS and other reasons are increasing injection rates and bandwidth.
  - “A preliminary evaluation of the hardware acceleration of the Cray Gemini Interconnect for PGAS languages and comparison with MPI” Shan et al.

A representative app : CTH

- Eulerian multi-material modeling application.
- 3-d, finite volume stencil computation.
- BSP with message aggregation (BSPMA).



Miniapps : Tools enabling exploration

Sandia National Laboratories

Focus	Proxy for a key app performance issue
Intent	Tool for codesign: output is information
Scope of change	Any and all
Size	A few thousand lines of code
Availability	Open source (LGPL)
Developer/owner	Application team
Life span	Until its no longer useful

Related:

Benchmark	Output: metric to be ranked.
Compact app	Application relevant answer.
Skeleton app	Inter-process comm, application “fake”
Proxy app	Über notion

Mantevo(.org) project

Miniapp	
CloverLeaf	Compressible Euler eqns, explicit 2nd order accurate
CoMD	Molecular dynamics (SPaSM)
HPCCG	Unstructured implicit finite element
miniFE	Implicit finite element solver
miniGhost	FDM/FVM explicit (halo exchange focus)
miniMD	Molecular dynamics (Lennard-Jones)
miniXyce	SPICE-style circuit simulator
mini"Aero"*	*In development*
Mini"Wave"*	*In development*
miniITC-FE	Implicit Thermal Conduction (Kokkos-based)
miniExDyn-FE	Explicit Dynamics (Kokkos-based)
phdMesh	Explicit FEM: contact detection

"Summary of Work for ASC L2 Milestone 4465: Characterize the Role of the Mini-Application in Predicting Key Performance Characteristics of Real Applications", R.F. Barrett (PI), P.S. Crozier, D.W. Doerfler (PM), S.D. Hammond, M.A. Heroux, H.K. Thornquist, T.G. Trucano, and C.T. Vaughan, Sandia Technical Report SAND 2012-4667, Sandia National Laboratories, 2012. *(Journal submission in preparation.)*

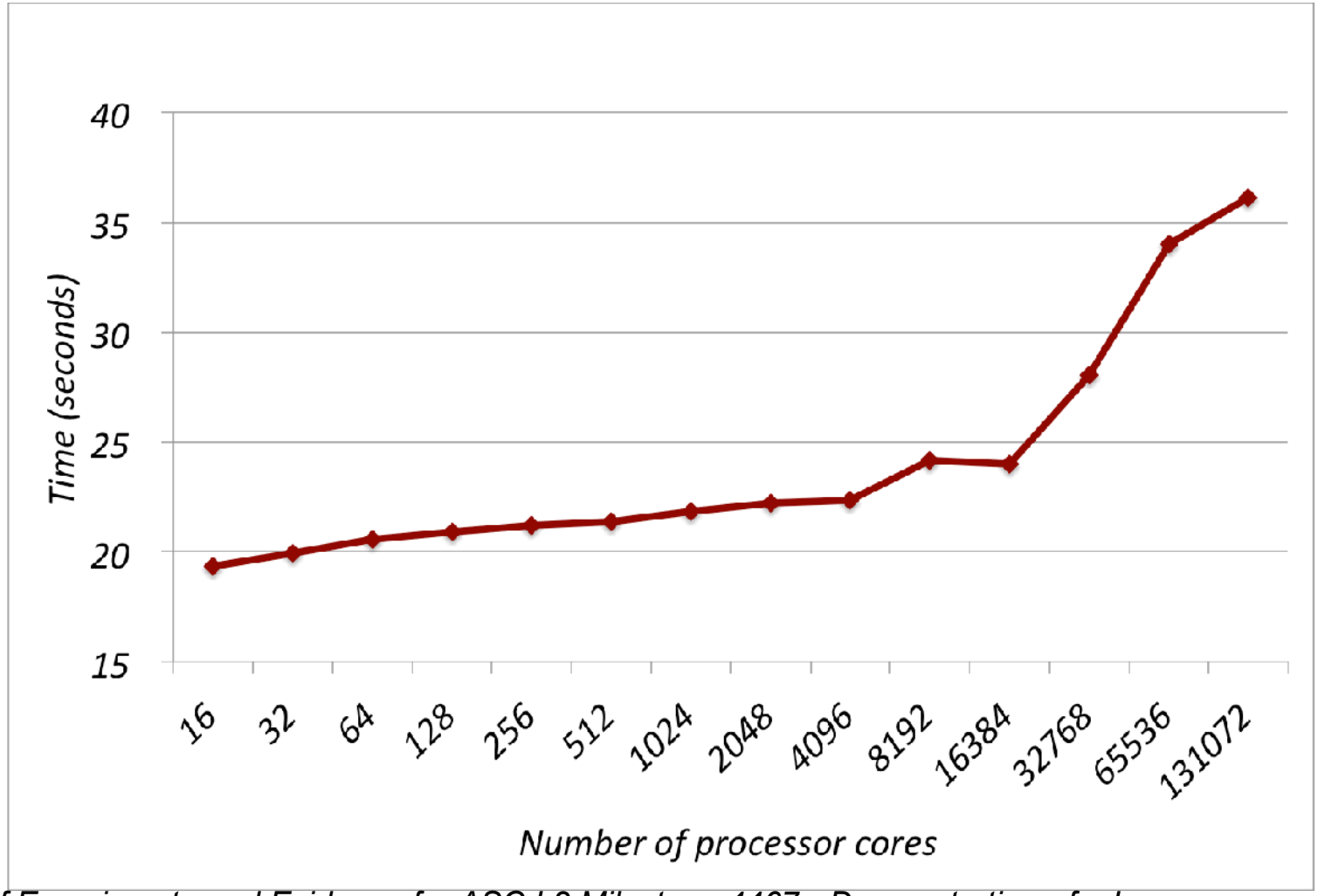
Computing resources

Sandia National Laboratories

	Cielo	Chama	Piz Daint
Processor	AMD Magny-Cours	Intel Sandy Bridge	Intel Sandy Bridge
Nodes	8,518	1,232	2,256
Sockets/node	2	2	2
Cores/socket	8	8	8
Total number of cores	136,288	19,712	36,096
Clock speed (GHz)	2.4	2.6	2.6
Memory/node (GB)	32	32	32
Memory	DDR4 1333 MHz	DDR3 1600 MHz	DDR3 1600 MHz
Socket connection	HyperTransport	PCIe Gen2/QPI	PCIe Gen3
Interconnect	Gemini 3D torus	Qlogic QDR fat tree	Cray Aries Dragonfly

miniGhost on Cielo

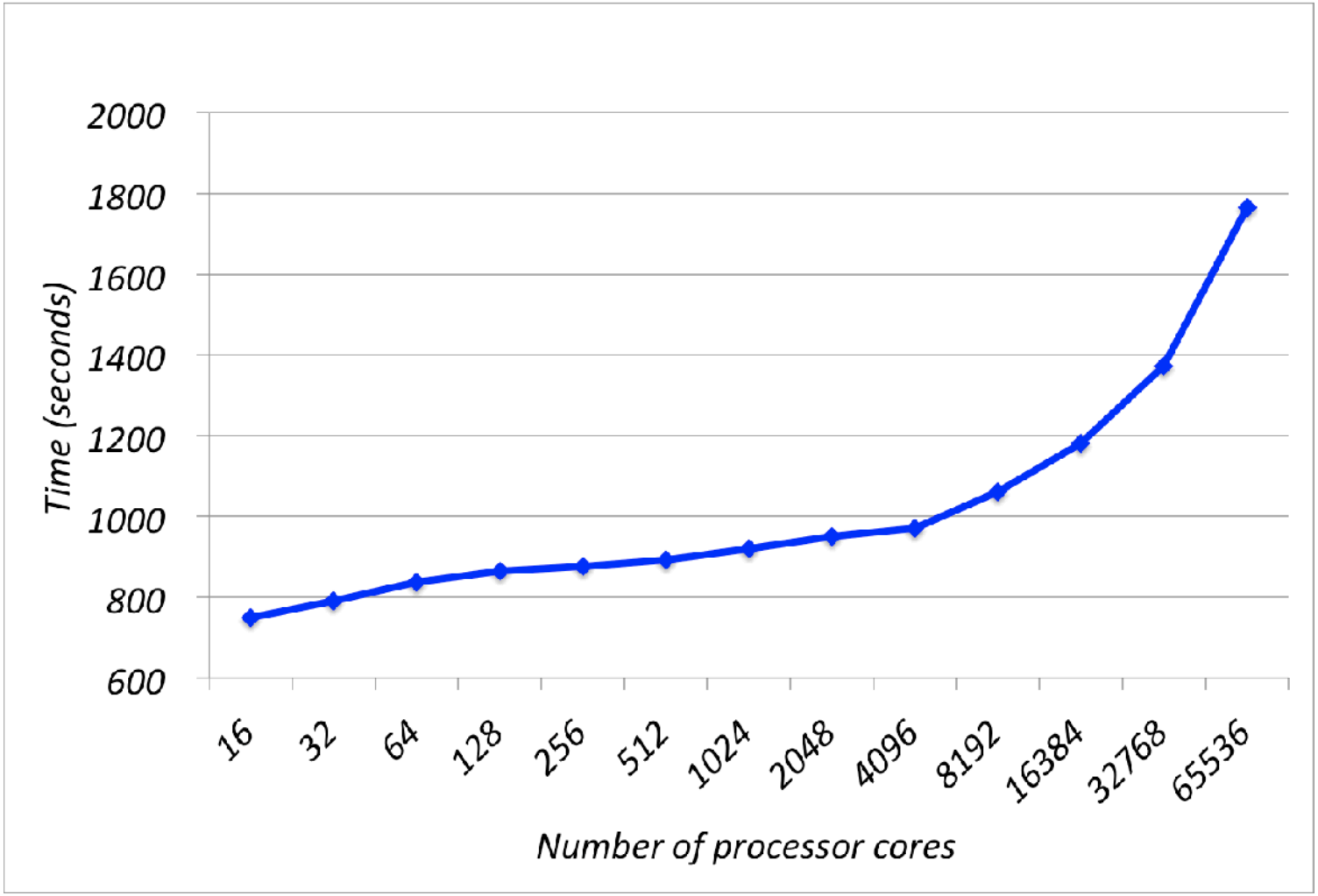
Nearest neigh comm in 3d code on 3d torus



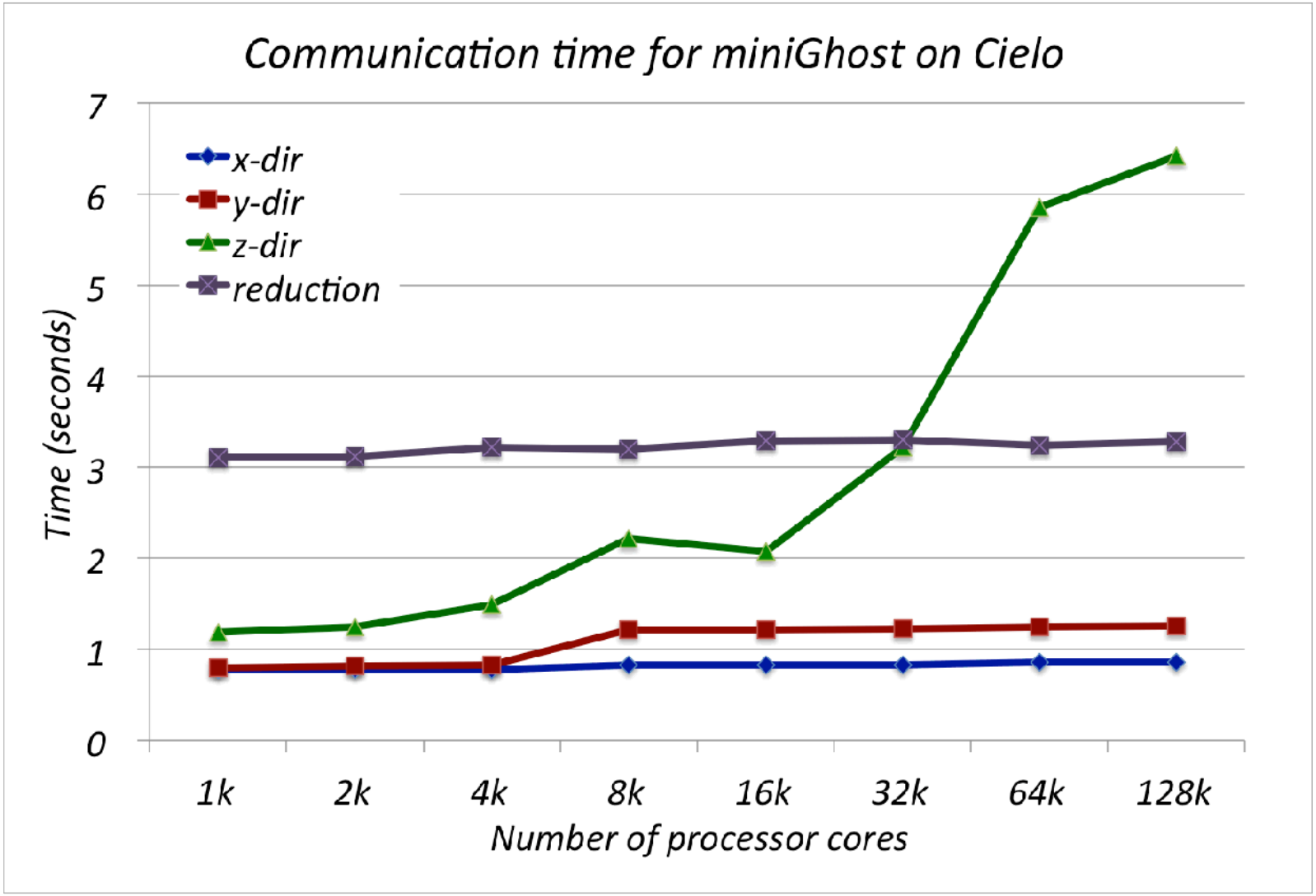
"Report of Experiments and Evidence for ASC L2 Milestone 4467 - Demonstration of a Legacy Application's Path to Exascale", S.M. Kelly et al, Sandia Technical Report 2012-1750, March 2012.

CTH (sl) on Cielo

Nearest neigh comm in 3d code on 3d torus

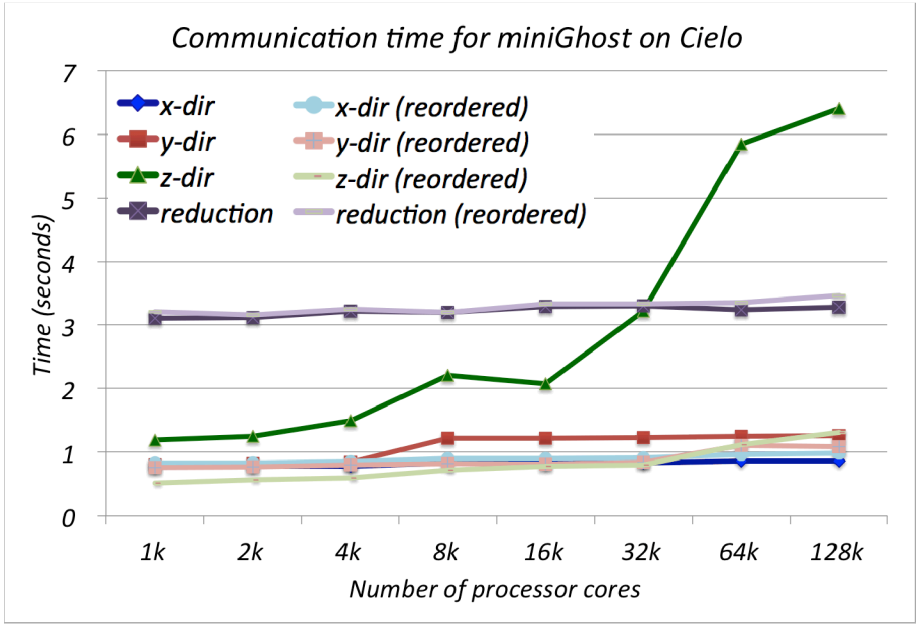


Dissecting communication time

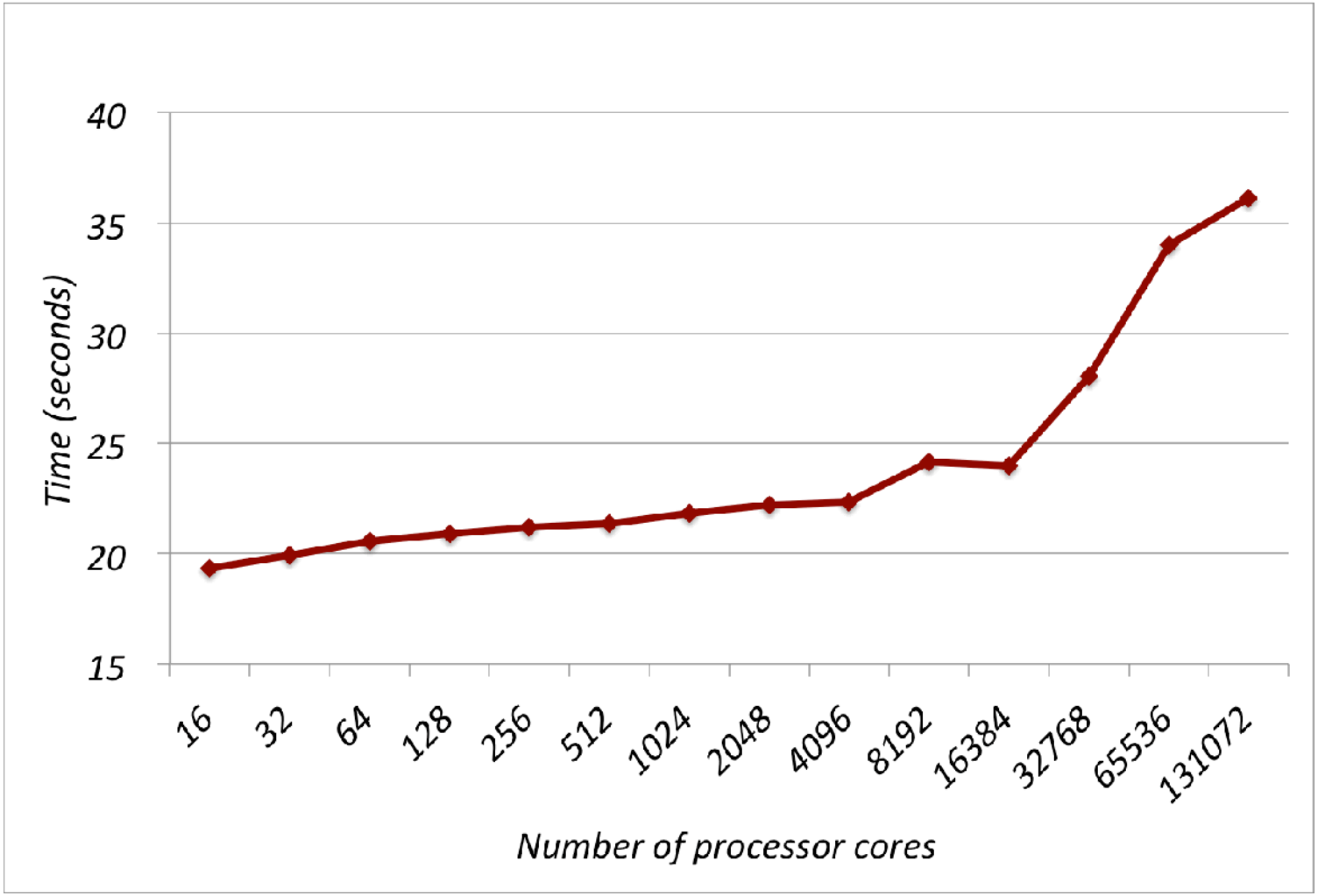


Managing interconnect hops miniGhost on Cielo

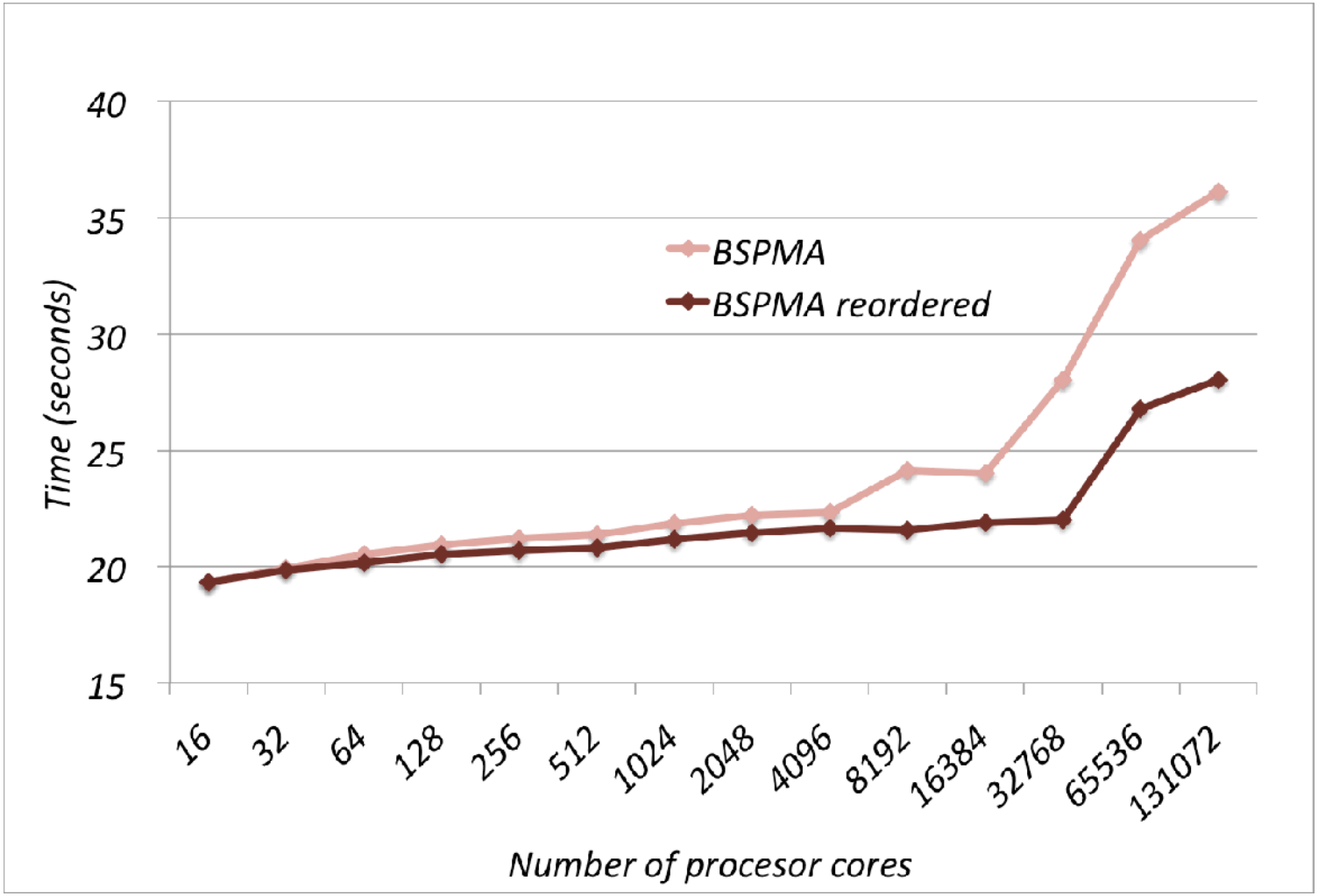
Number of MPI ranks	Regular Order X	Y	Z
16	0.0	0.0	0.0
32	0.0	0.0	0.0
64	0.0	0.0	0.3
128	0.0	0.0	1.0
256	0.0	0.0	1.0
512	0.0	0.1	2.0
1024	0.0	0.3	2.1
2048	0.0	0.3	2.7
4096	0.0	0.3	3.7
8192	0.0	0.5	5.1
16384	0.0	0.5	4.9
32768	0.0	0.5	5.6
65536	0.0	1.1	10.2
131072	0.0	1.1	10.1



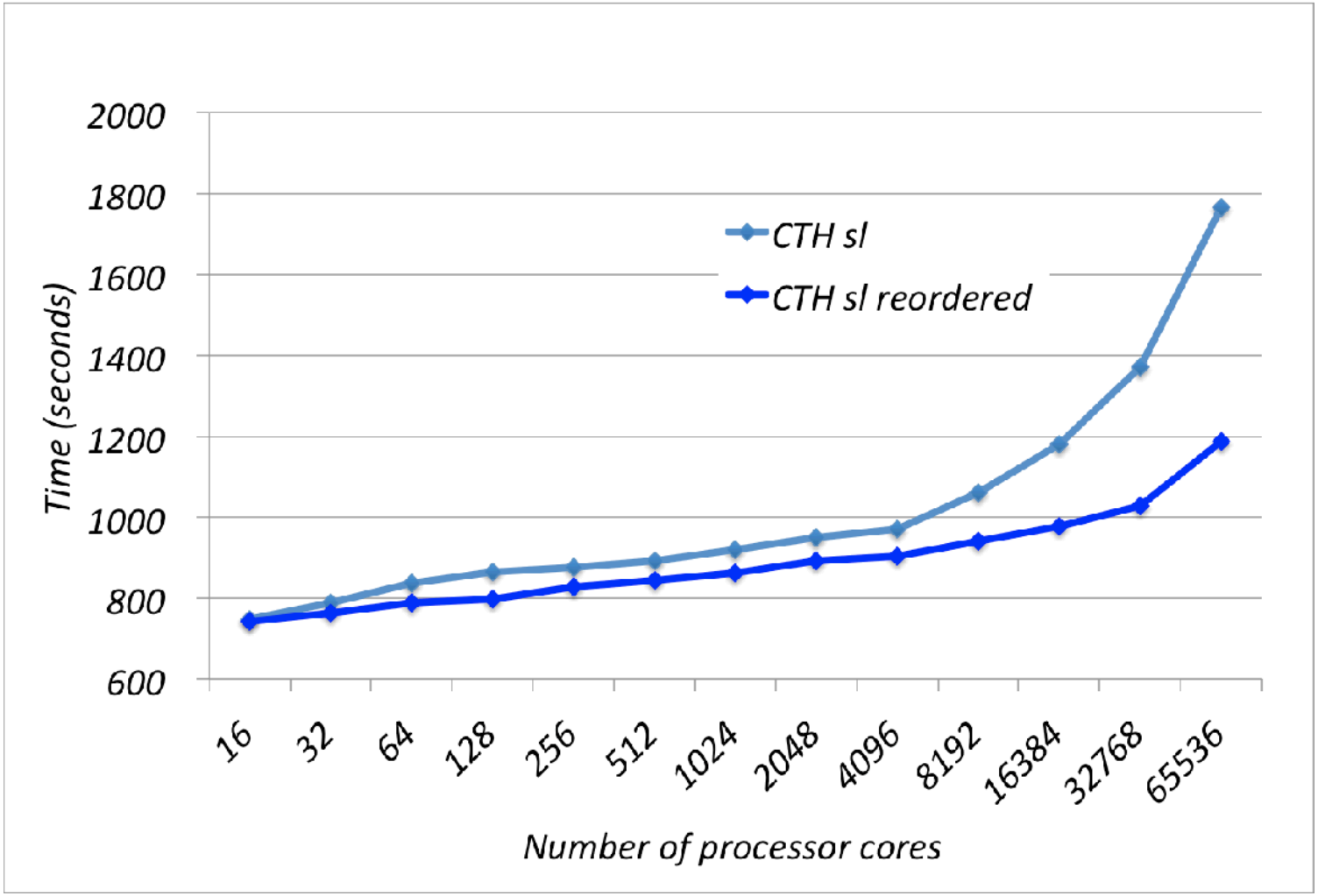
miniGhost on Cielo



miniGhost on Cielo



CTH on Cielo


2000
1800
1600
1400
1200
1000
800
600
Time (seconds)
CTH sl
CTH sl reordered
16
32
64
128
256
512
1024
2048
4096
8192
16384
32768
65536
Number of processor cores

Evidence for cause

- miniGhost hops in z-direction reduced, time reduced.
- CTH scaling improved, time reduced.
- CTH computation & global reduction time stays flat.
- Gemini counters: stalls.

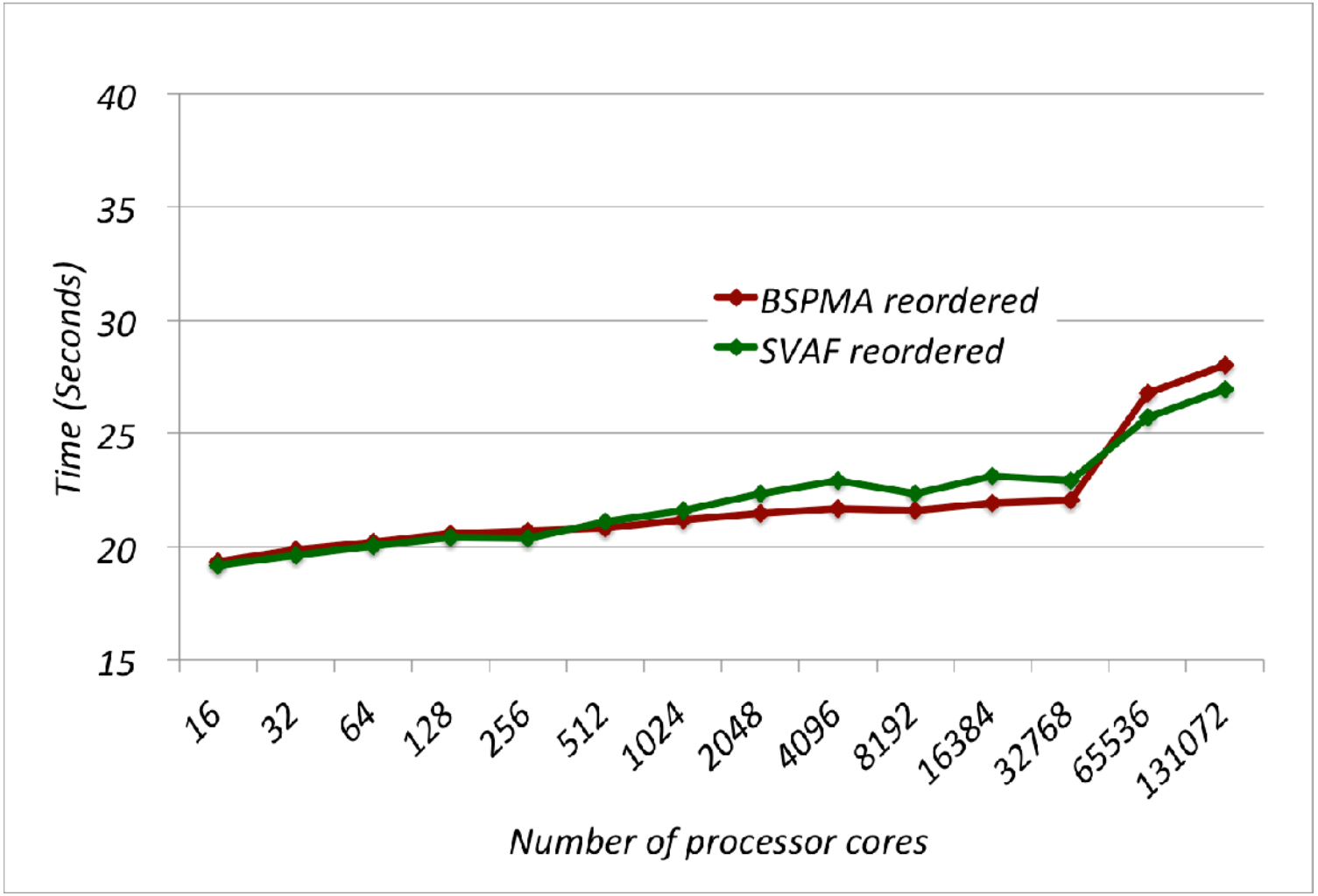
Alternative inter-node strategy : Single Variable Aggregated Faces (SVAF)

Sandia National Laboratories

```
do i = 1, num_tsteps

    do j = 1, num_vars
```

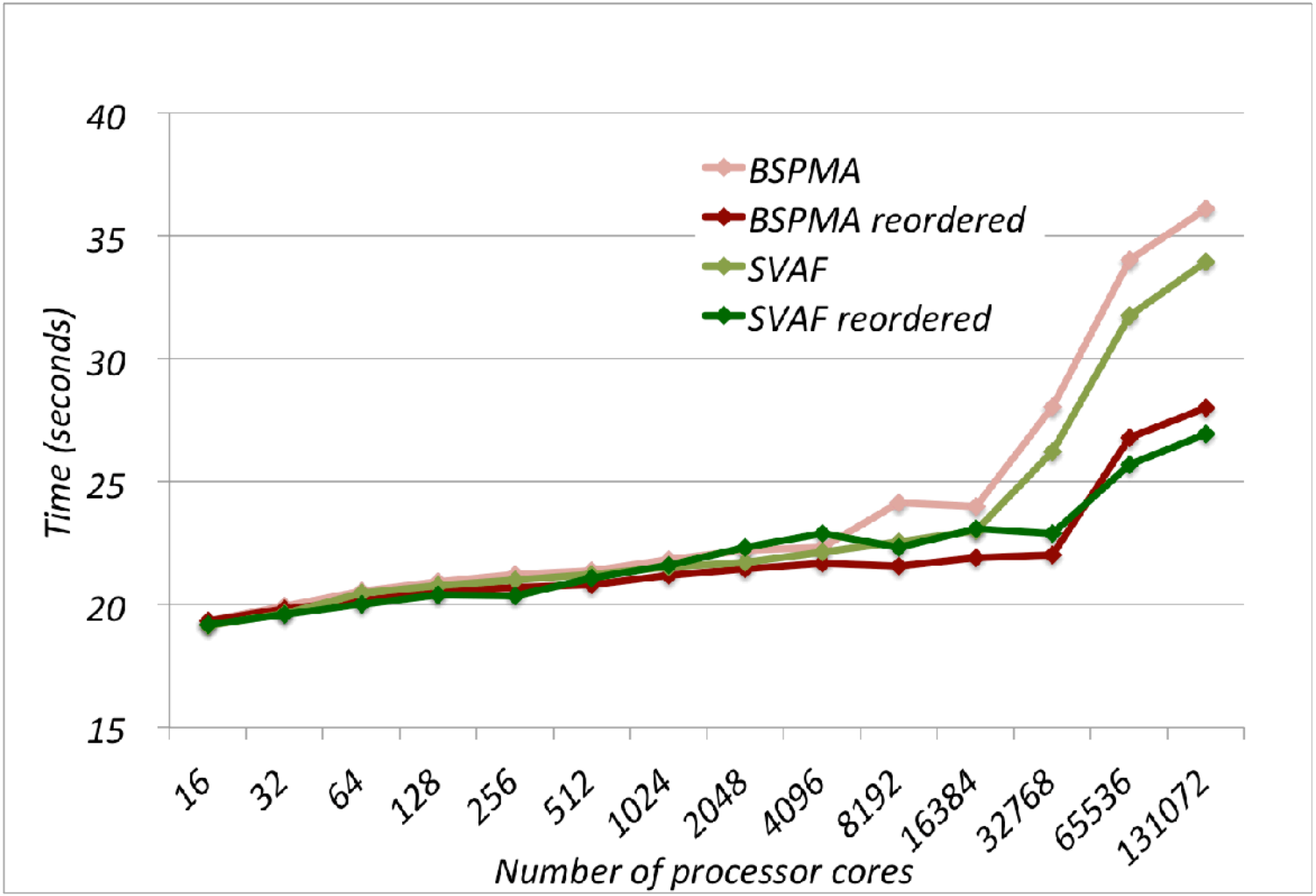
```
        compute

    end do

end do
```

miniGhost on Cielo* : SVAF and BSPMA



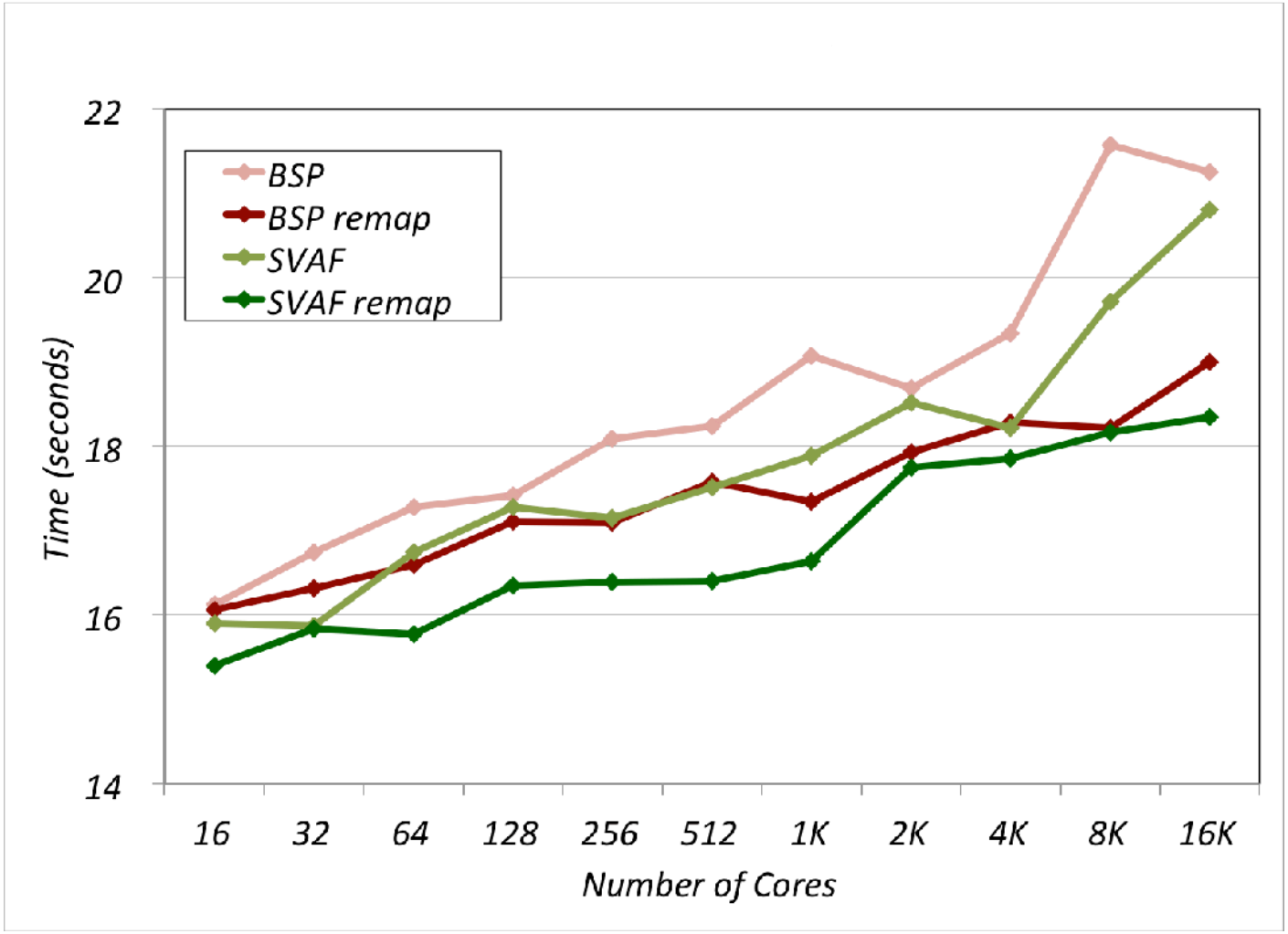
* Dedicated resource

miniGhost on Cielo* : SVAF and BSPMA



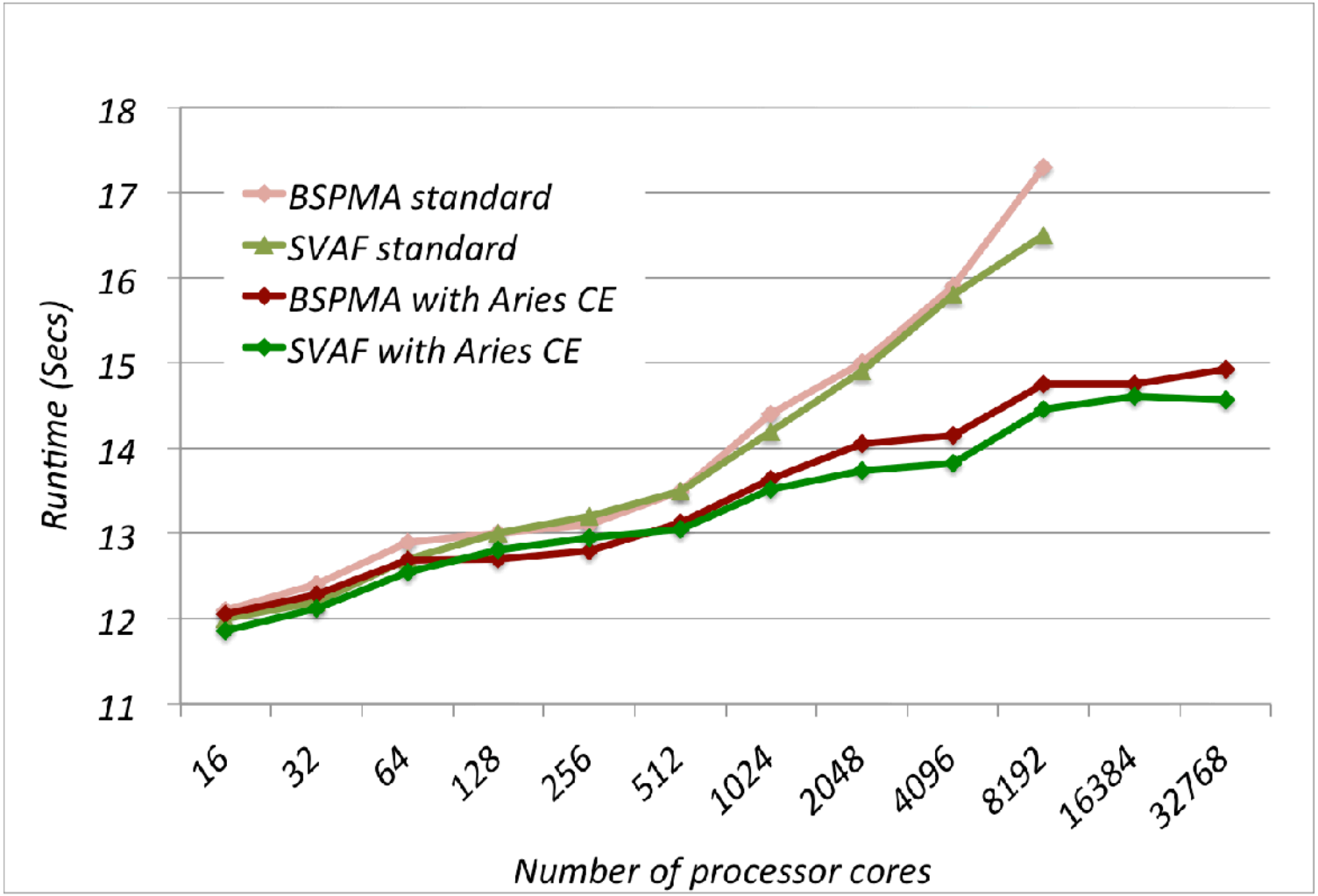
* Dedicated resource

miniGhost on Chama*: SVAF and BSPMA



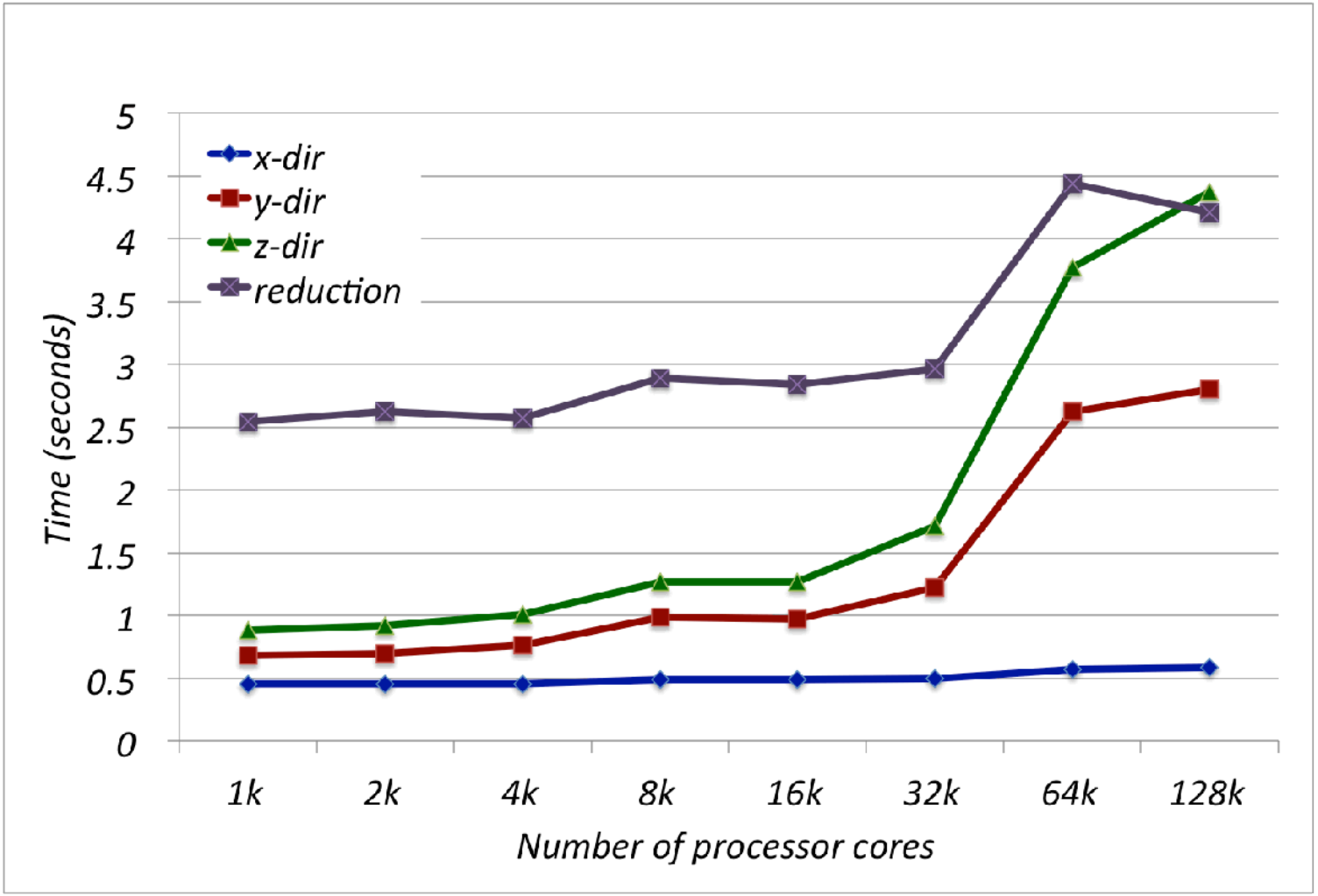
* Shared resource

miniGhost on Daint*: SVAF and BSPMA

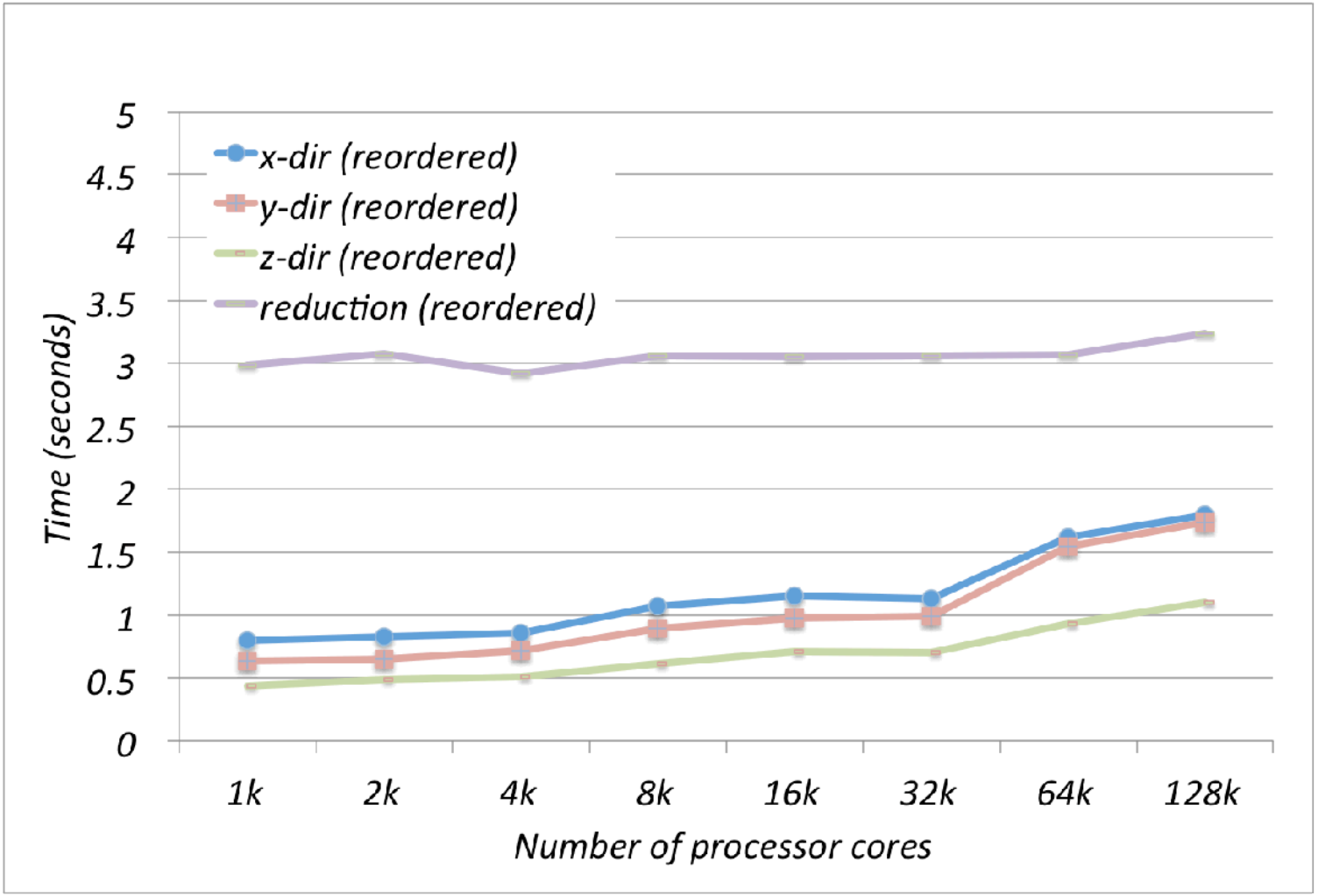


* Dedicated resource

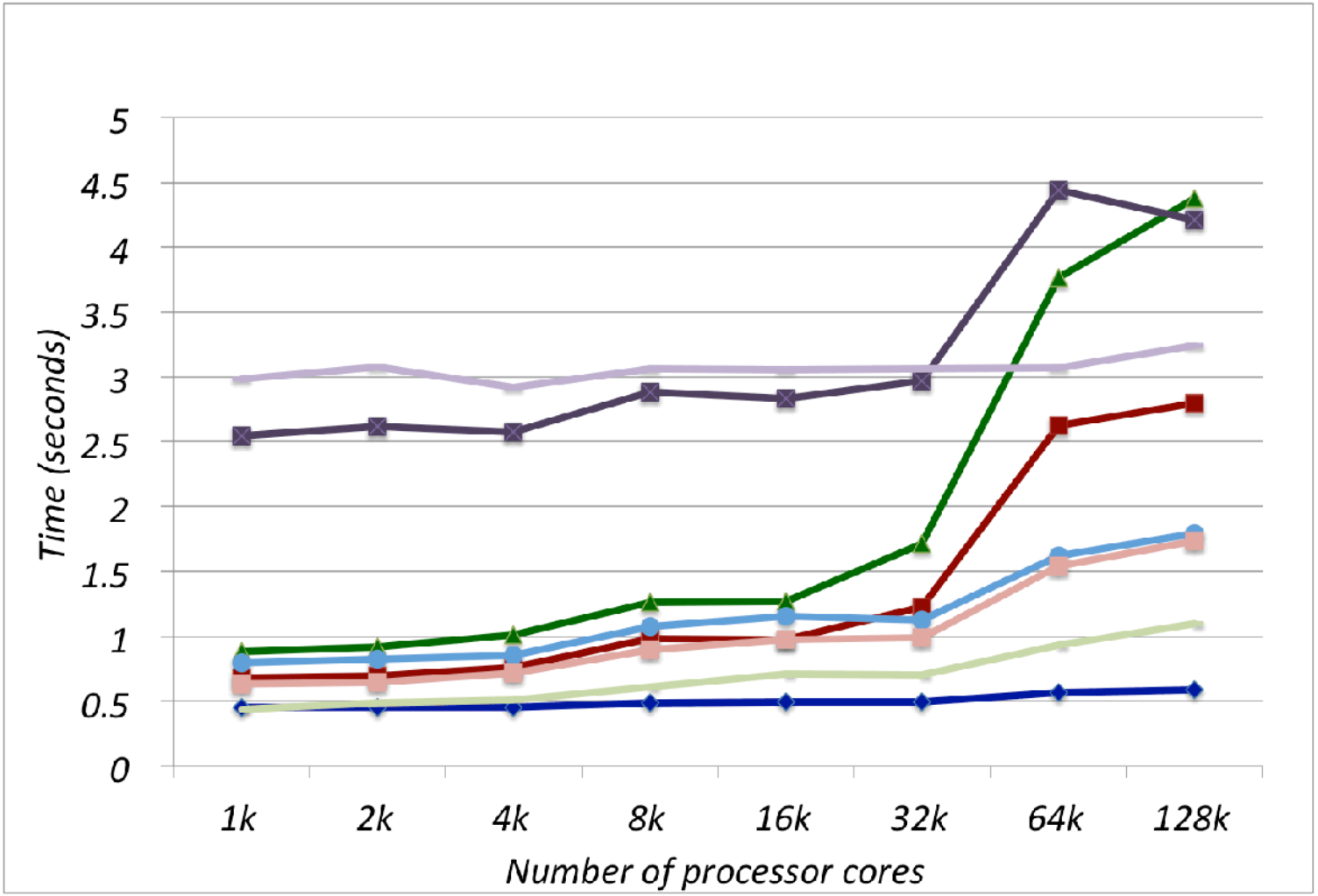
Dissecting communication time: miniGhost SVAF



Dissecting communication time: miniGhost SVAF



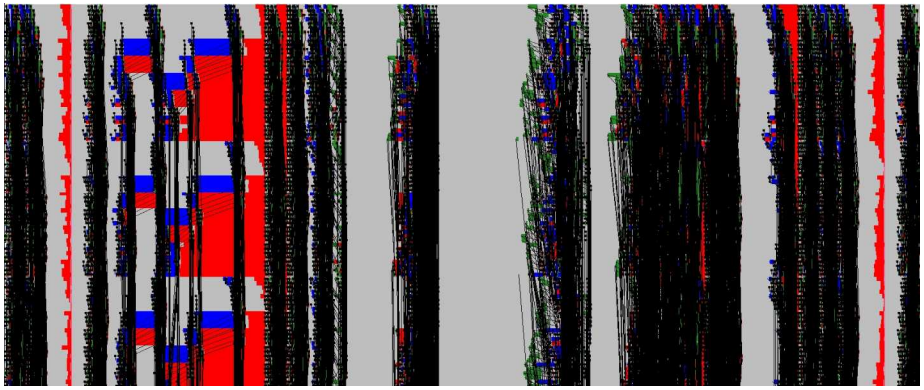
Dissecting communication time: miniGhost SVAF



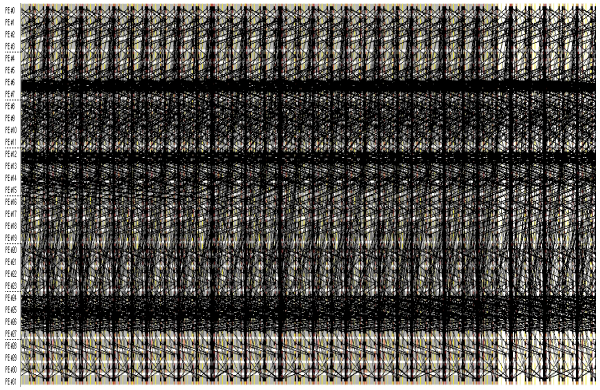
MPI + OpenMP

- Fewer larger messages
- We see similar scaling issues
- Not surprising: Its not the ability to move onto the network, it's the large volume on it.
- Still seeing MPI everywhere out-performing MPI+OpenMP.

Space-time profile



CTH

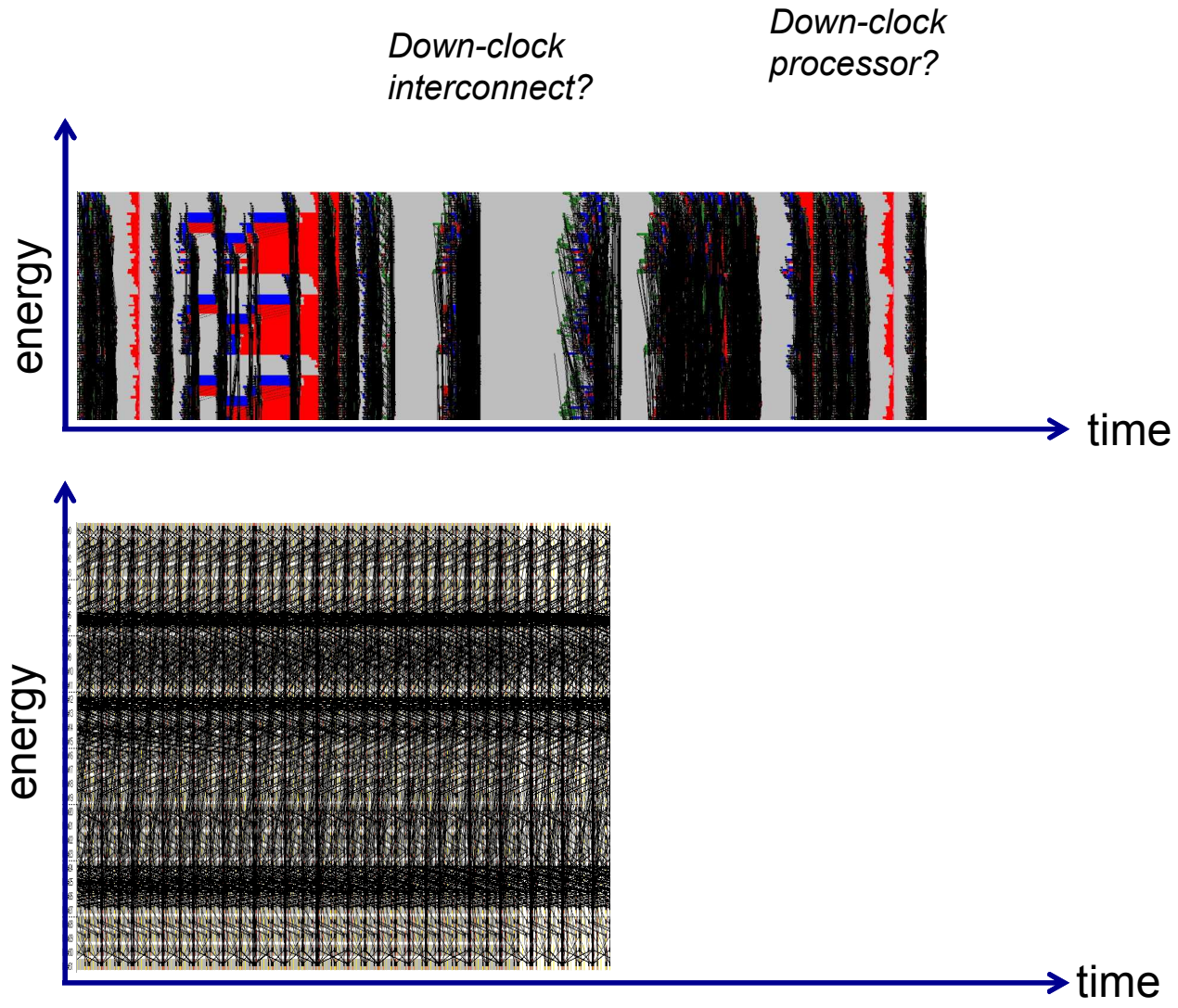


Should we try for this*?

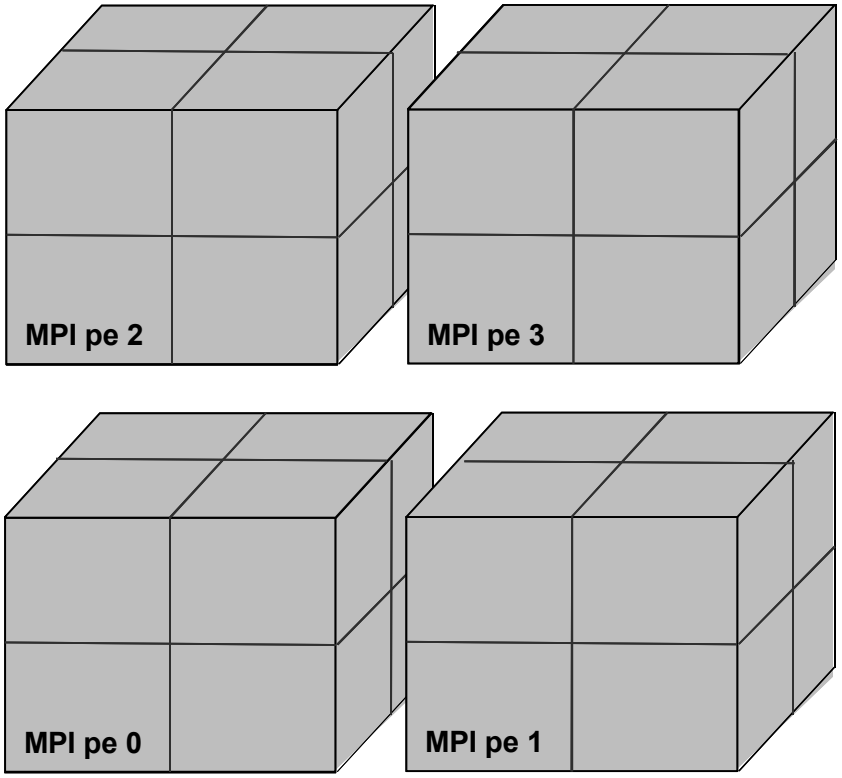
*not (yet?) CTH

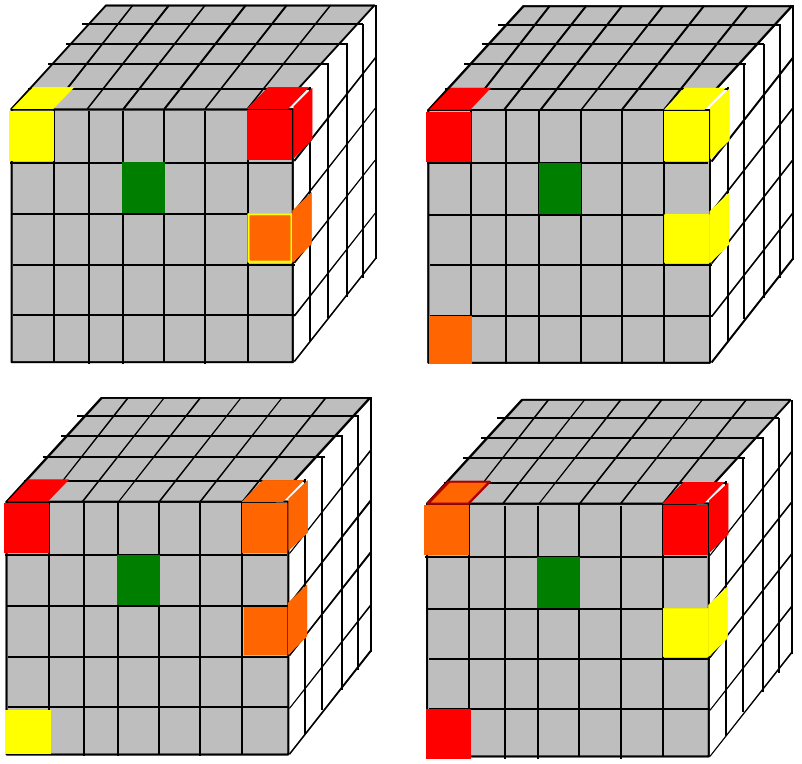
Computation: gray Communication: black Sync: red

Power perspective?



miniGhost tp





Data parallel
thread

Task parallel: some representative task workloads
Computation
Computation + MPI
Computation + BC
Computation + BC + MPI

miniGhost tp : Adding AMR

Sandia National Laboratories

- Diffusion over the 3d domain with random initial conditions and reflective boundary conditions.

  Two options:

  - uniform refinement, or
  - refinement based on the boundary or volume of an object being moved through the mesh and changing size. So, for example, a shock front can be simulated by refining based on a sphere which starts small and grows in size as the problem advances.
- Refinement within blocks.
- A block is refined into 8 blocks.
- Neighbors must be within one level of refinement.
- Computation is self-contained within a block.
- Communication aggregated to BSP model.
  - Excellent candidate for task parallelism version.

Summary

Sandia National Laboratories

- Concrete steps can be taken to preparing applications for what we see impacting exascale computation.
- These steps lead to stronger performance on current and emerging architectures.
- Interconnect capabilities changing in positive ways, but
- Process placement still matters.

On-going work

- Intel ϕ, Kepler, Calxeda, ARM, Tilera, Convey, …
- Revolutionary programming models, languages, mechanisms

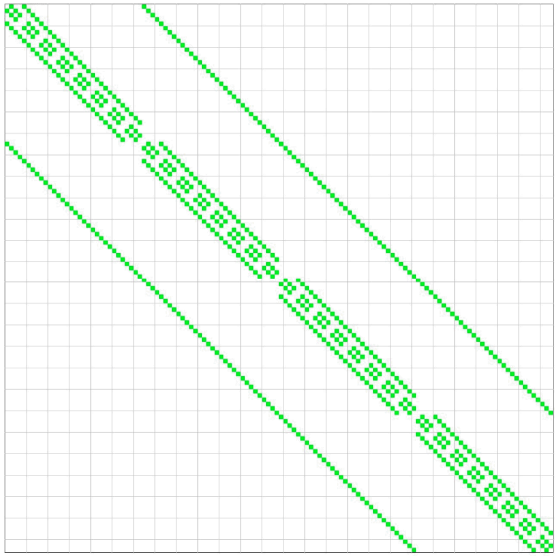
More information

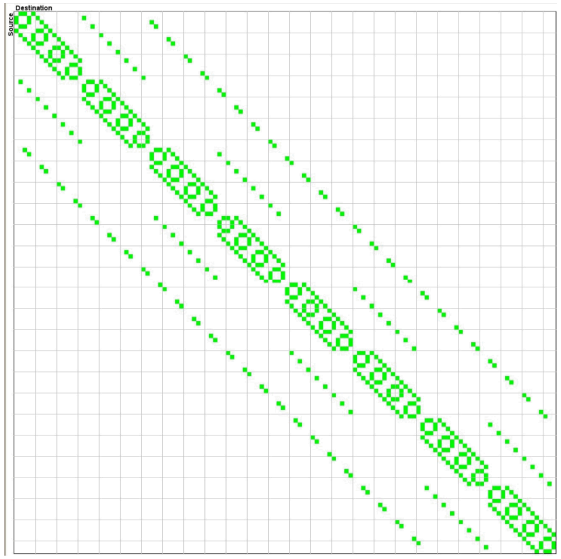
- Mantevo : http://mantevo.org
- *MiniGhost: A Miniapp for Exploring Boundary Exchange Strategies Using Stencil Computations in Scientific Parallel Computing; Version 1.0,* R.F. Barrett, C.T. Vaughan, and M.A. Heroux, Technical Report SAND-2012-10431, Sandia National Laboratories, 2012. This is a "living document", previously reported as SAND-2012-2437.
- *Summary of Work for ASC L2 Milestone 4465: Characterize the Role of the Mini-Application in Predicting Key Performance Characteristics of Real Applications"*, R.F. Barrett (PI), P.S. Crozier, D.W. Doerfler (PM), S.D. Hammond, M.A. Heroux, H.K. Thornquist, T.G. Trucano, and C.T. Vaughan, Sandia Technical Report SAND 2012-4667, Sandia National Laboratories, 2012.
- *On the Role of Co-design in High Performance Computing*, R.F. Barrett, S. Borkar, S.S. Dosanjh, S.D. Hammond, M.A. Heroux, X.S. Hu, J.Luitjens, S.Parker, J. Shalf, and L.Tang, 2013, Under review.

Acknowledgements

- CSCS (Swiss Computing Center) for access to and support of Piz Daint, a Cray XC; discussions with Sadaf Alam.
- ACES (Los Alamos / Sandia computing center), funded by Department of Energy's National Nuclear Security Administration (NNSA) Advanced Simulation and Computing (ASC) campaign for the use of Cielo in dedicated mode.
- Thanks to Darren

Additional slides




Destination
Source