



Argonne
NATIONAL
LABORATORY

... for a brighter future



SciDAC Institute for Ultrascale Visualization

www.ultravis.org



U.S. Department
of Energy

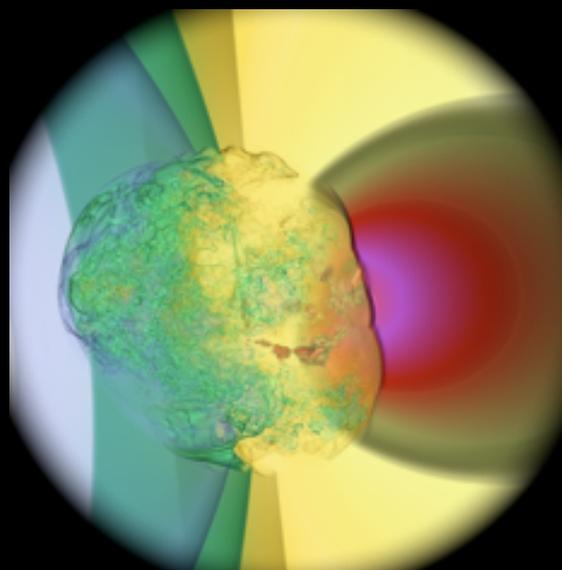
UChicago ►
Argonne_{LLC}



U.S. DEPARTMENT OF ENERGY

A U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC

Scalable Approaches to Analysis of Scientific Data



Volume rendering of x-velocity in time-step 1530
of a hydrodynamics simulation of a core-collapse
supernova.

Tom Peterka

tpeterka@mcs.anl.gov

Mathematics and Computer Science Division

Top 5 Strange Things Heard at a Conference

1. Large scale parallel visualization on HPC machines

2. Exascale machines will use wireless interconnects

3. Edge computing

4. Speedup and scalability on a single core

5. Pedaflops and other misspellings

Parallel Analysis and Visualization on HPC Architecture

Two questions arise.

Why parallel?

Architectural limitations on power, clock speed, cooling
Memory size
Parallelism is the *ONLY* way to continued improvements
There are many kinds of parallelism
You are using it already, whether you realize it or not
Multiple architectures can benefit
Room for new research

Why supercomputers?

That's where science is computed and data reside

A Growing Rift

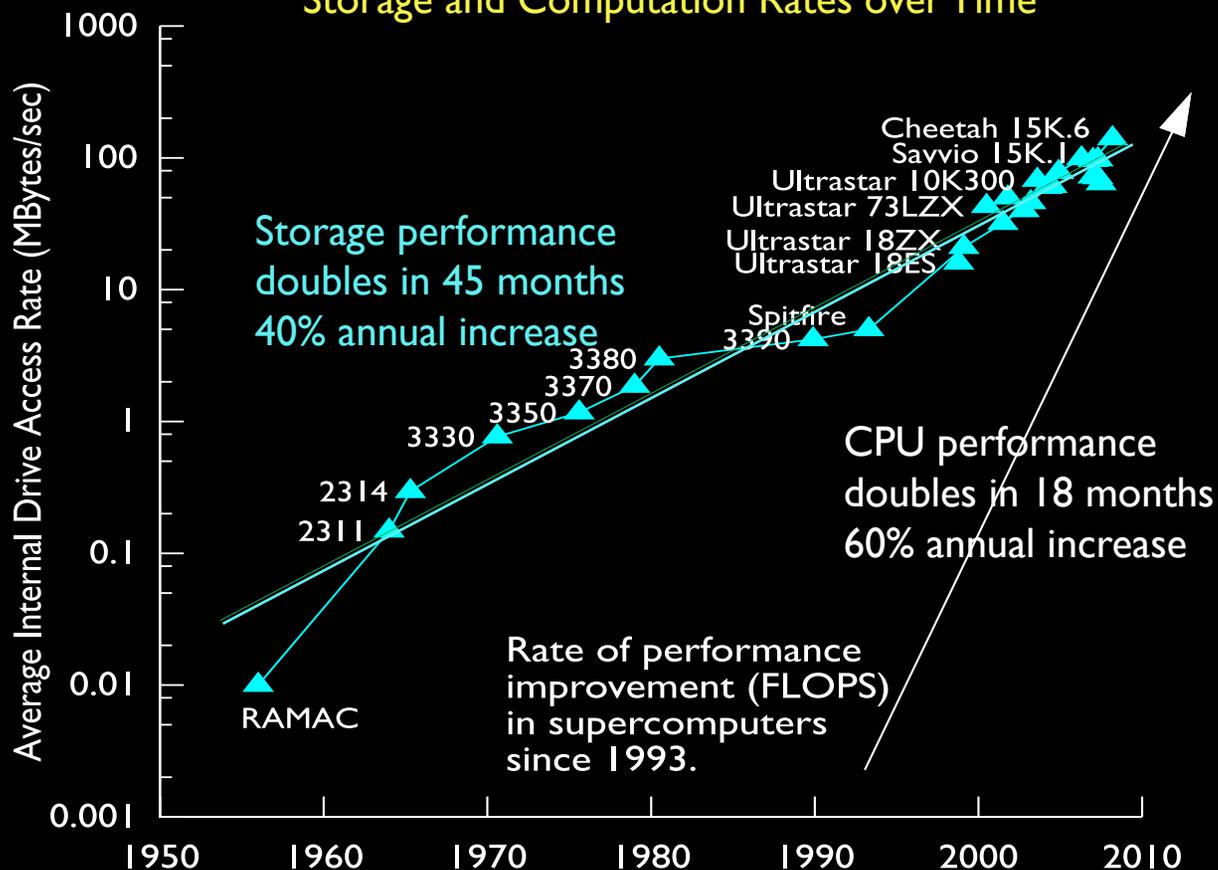
We are computing more data, faster than we can manage.

Total data of selected 2008 INCITE awards as of 6/08

Domain	Data size (TB)	PI
Astrophysics	375	Lamb
Climate	355	Washington
Materials	105	Wolverton
Fusion	54	Klasky

Ref: Private correspondence Katherine Riley, Argonne ALCF, 2008.

Storage and Computation Rates over Time



Ref: Rob Ross, Visualization and Parallel I/O at Extreme Scale, SciDAC '08

More than Peak FLOPS

More than any other factor, disk I/O rate limits analysis capability.

Normalized Storage / Compute Metrics

Machine	Storage B/W (GB/s)	Storage Size (PB)	FLOPS (Pflop/s)	Flops per byte stored
LLNL BG/L	43	2	0.6	$O(10^4)$
Jaguar XT4	42	0.6	0.3	$O(10^4)$
Intrepid BG/P	50	5	0.6	$O(10^4)$
Roadrunner	50	5	1.0	$O(10^5)$
Jaguar XT5	42	5	1.4	$O(10^5)$

-The average flops per byte of parallel I/O disk access today is between 10,000 and 100,000

-In 2001, this number was approximately 500. Ref: John May, 2001.

-DOE science applications generate results at an average rate of 40 flops per byte of data. Ref: Murphy et al. ICS'05.

-Applications can only afford to save between 1-10% of what they compute.
-With postprocessing, what is not saved cannot be analyzed.

Percent Saved of Computed Data

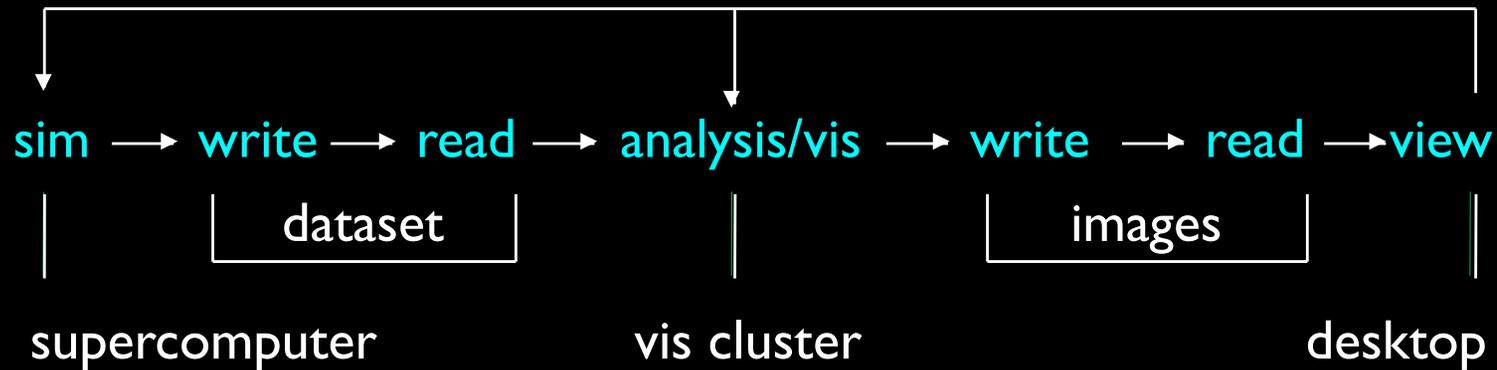
Code	Domain	% Saved	PI
FLASH	Astrophysics	10	Ricker
Nek5000	CFD	1	Fischer
CCSM	Climate	1	Jacob
GCRM	Climate	10	Cram
S3D	Combustion	1-5	Bennett

Ref: CScADS Scientific Data Analysis & Visualization Workshop '09

Our Science Workflow Cannot Scale Indefinitely.

“Life is not fun when you’re banging your head against a brick wall all the time.”

– John McEnroe, winner of 7 grand slam tennis titles



"Models that can currently be run on typical supercomputing platforms produce data in amounts that make storage expensive, movement cumbersome, visualization difficult, and detailed analysis impossible. The result is a significantly reduced scientific return from the nation's largest computational efforts." -Mark Rast, Laboratory for Atmospheric and Space Physics, University of Colorado

A Solution

Move some of the visualization to the data.

The increasing demands for analysis and visualization can be met by performing more analysis and **visualization tasks directly on supercomputers** traditionally reserved for simulation.

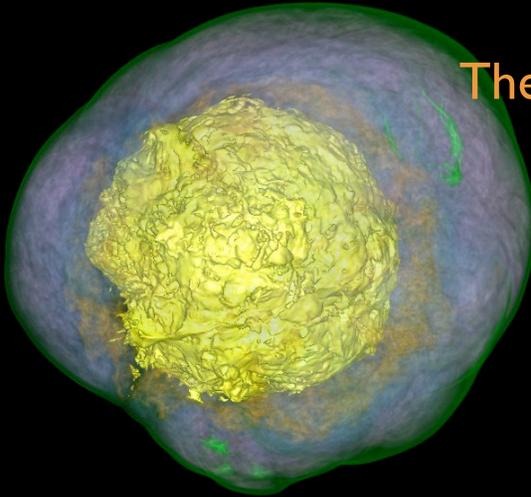
Potential benefits: **Increased overall performance, reduced cost, tighter integration** of analysis and visualization in computational science.

Potential drawbacks: **Reduced per-core performance, increased load on computing resources, potential to crash computations.**

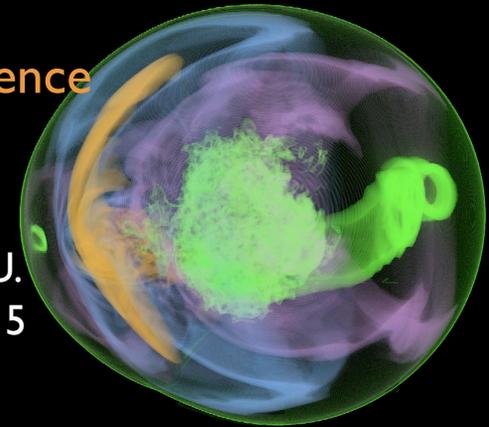
Applications

The science behind the computer science

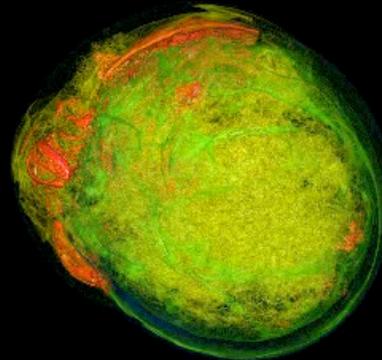
Volume rendering of shock wave formation in core-collapse supernova dataset, courtesy of John Blondin, NCSU. Structured grid of 1120^3 data elements, 5 variables per cell.



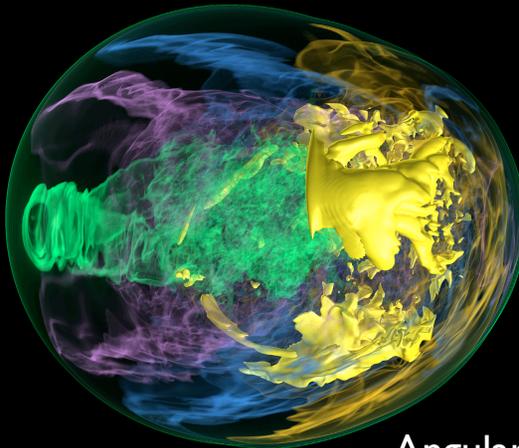
Pressure at time-step 1530



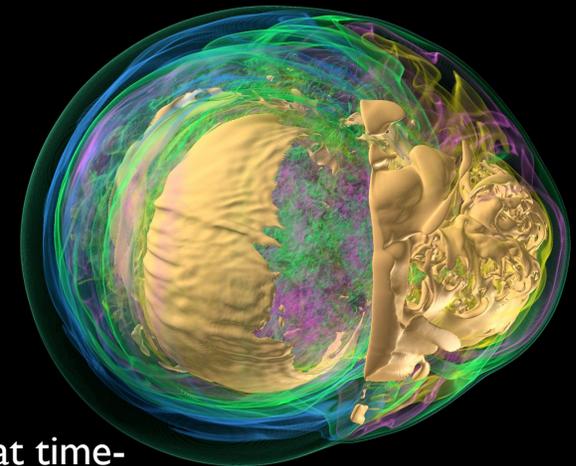
Angular momentum at time-step 1403



Entropy over 100 time-steps



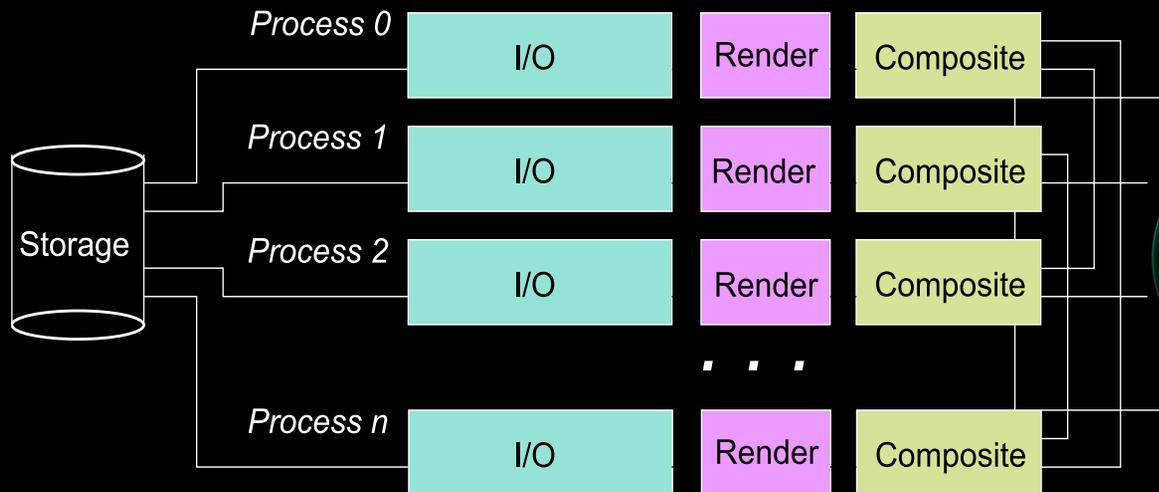
Angular momentum at time-step 1492



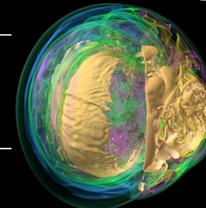
Entropy at time-step 1518

Algorithms

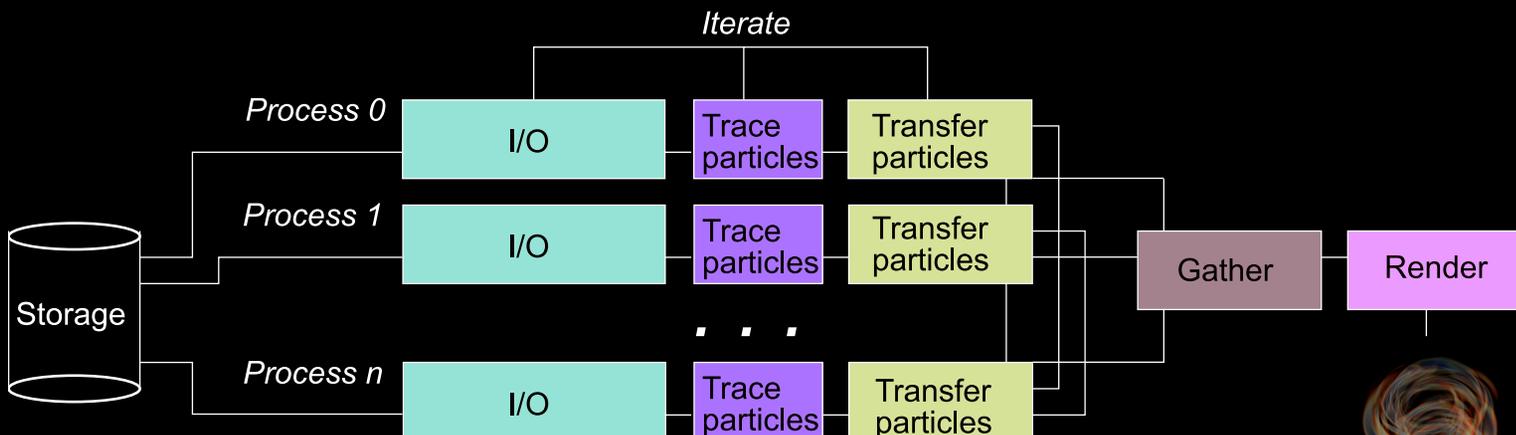
Scalar and vector fields



Parallel structure for volume rendering algorithm consists of 3 stages performed in parallel



Parallel Volume Rendering on the IBM Blue Gene/P. EGPGV'08.



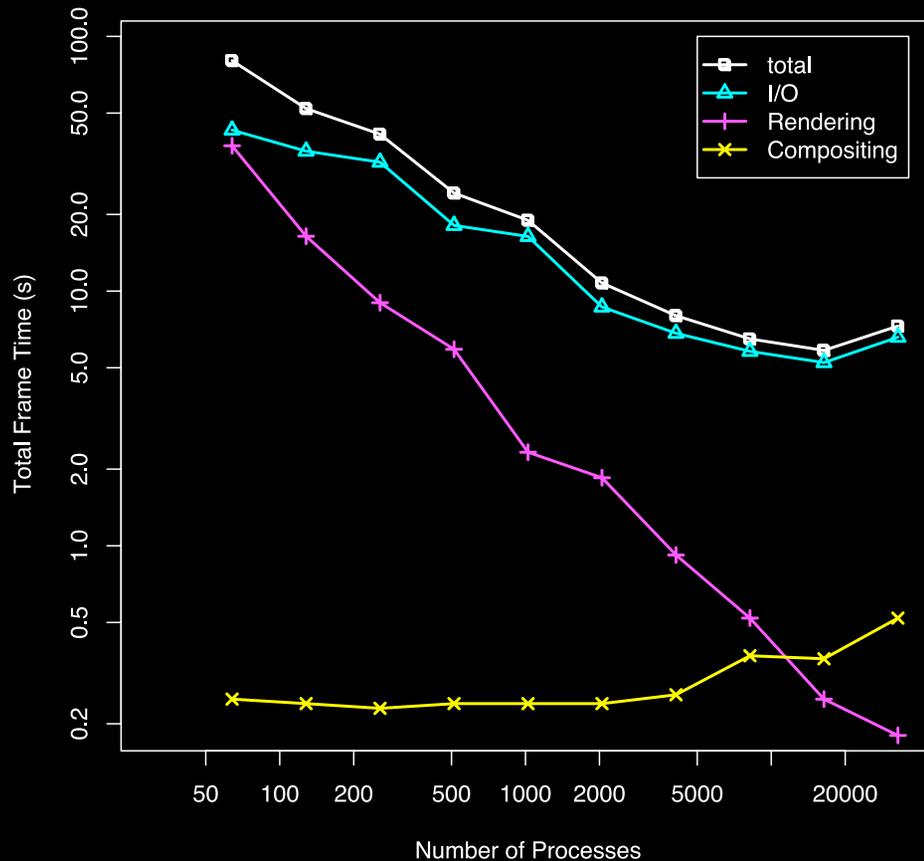
Parallel structure for flow visualization algorithm consists of iterations of particle tracing and transfer, followed by a rendering stage.



Performance

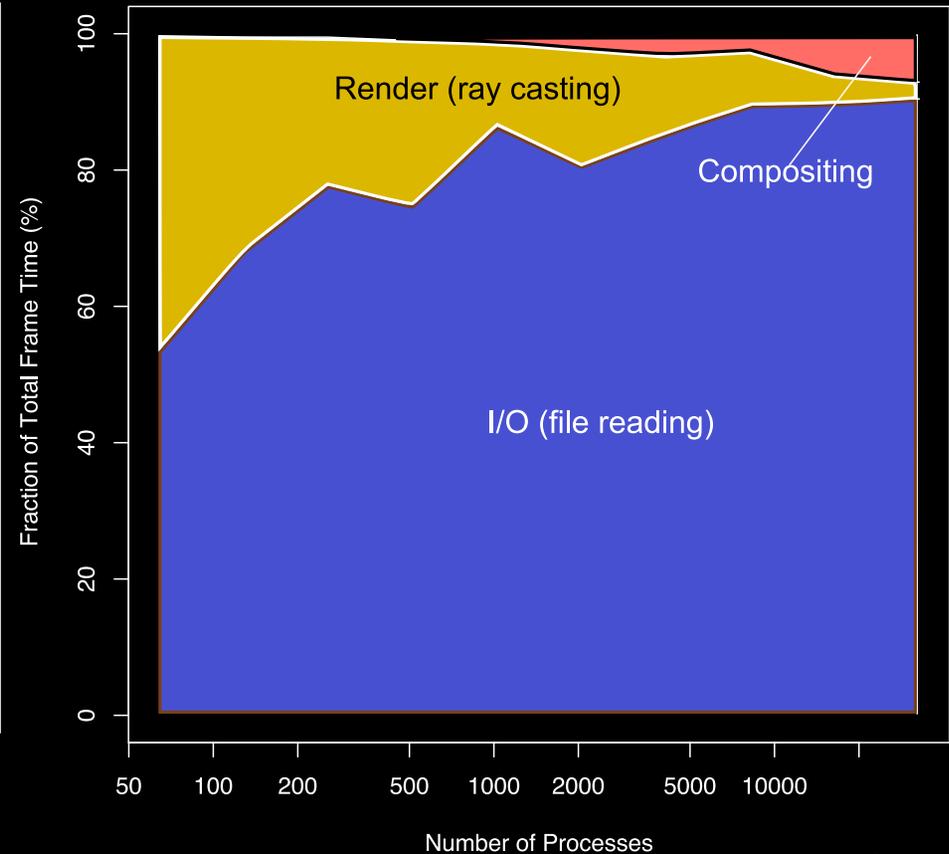
Total and component time

Total and Component Time



Total frame time and individual component times. Raw data format, 1120^3 , image size 1600^2 .

Time Distribution

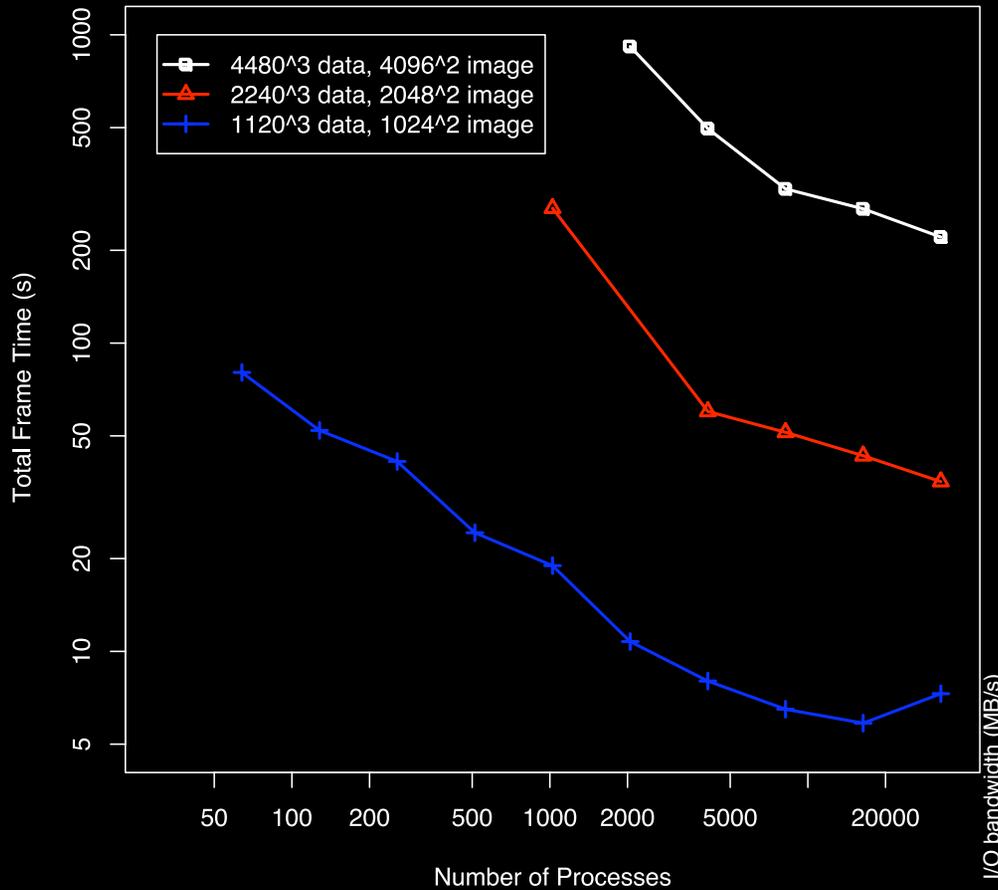


The relative percentage of time in the stages of volume rendering as a function of system size. Large visualization is primarily dominated by data movement: I/O and communication.

Performance

Large-scale results

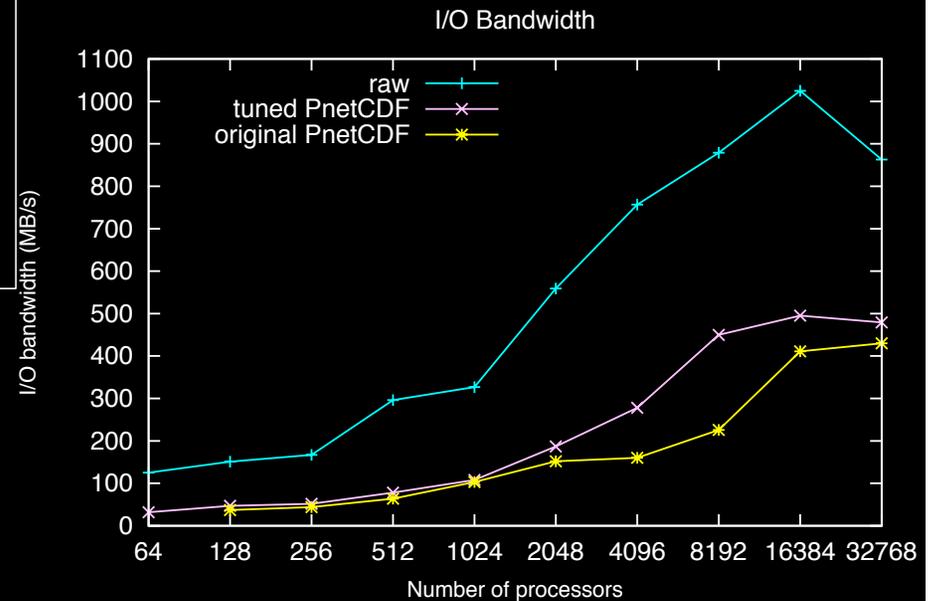
Volume Rendering End-to-End Performance



Scalability over a variety of data, image, and system sizes. A number of performance points exist for each data size.

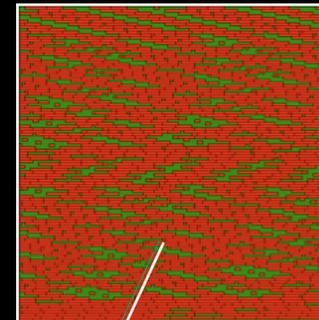
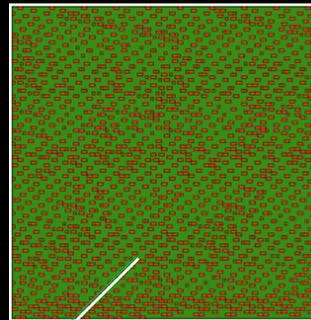
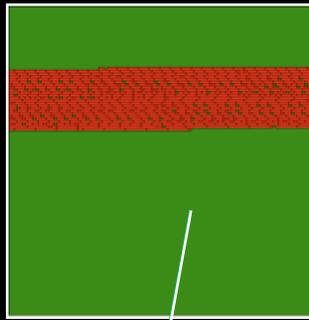
Grid Size	Time-step size (GB)	Image size (px)	# Procs	Tot. time (s)	% I/O	Read B/W (GB/s)
2240 ³	42	2048 ³	8K	51	96	0.9
			16K	43	97	1.0
			32K	35	96	1.3
4480 ³	335	4096 ³	8K	316	96	1.1
			16K	272	97	1.3
			32K	220	96	1.6

Volume rendering performance at large size is dominated by I/O. While overall performance is scalable, I/O bandwidth is far below peak.

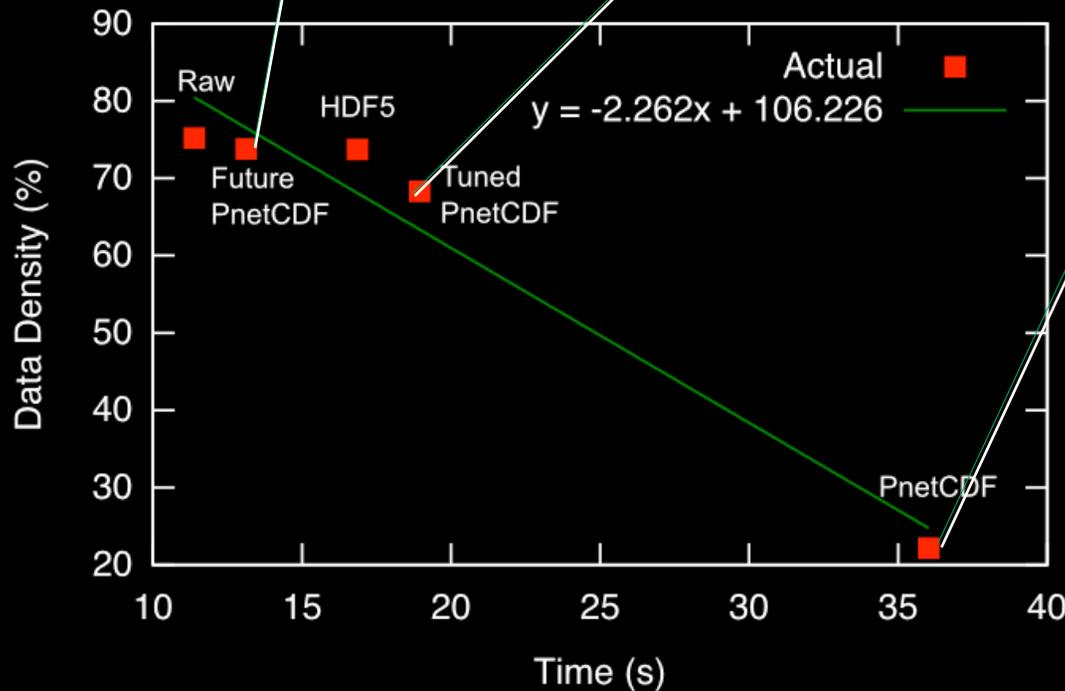


High-level File Organization

Reorganizing can produce drastic speedups.



I/O Mode Comparison



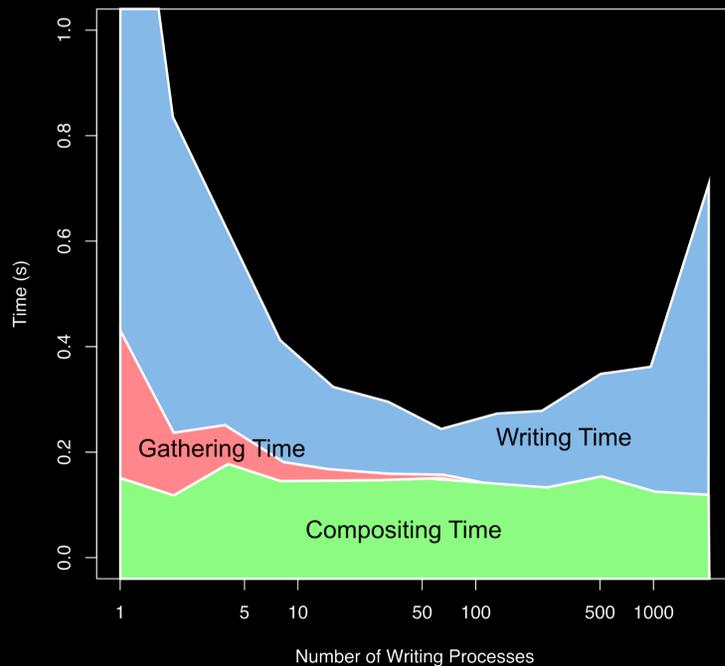
Changing data file layout can improve I/O performance. Top, different layouts produce improved file access patterns. Bottom, benchmarks confirm improved performance

Assessing Improvements to the Parallel Volume Rendering Pipeline at Large Scale. SC08 Ultrascale Visualization Workshop.

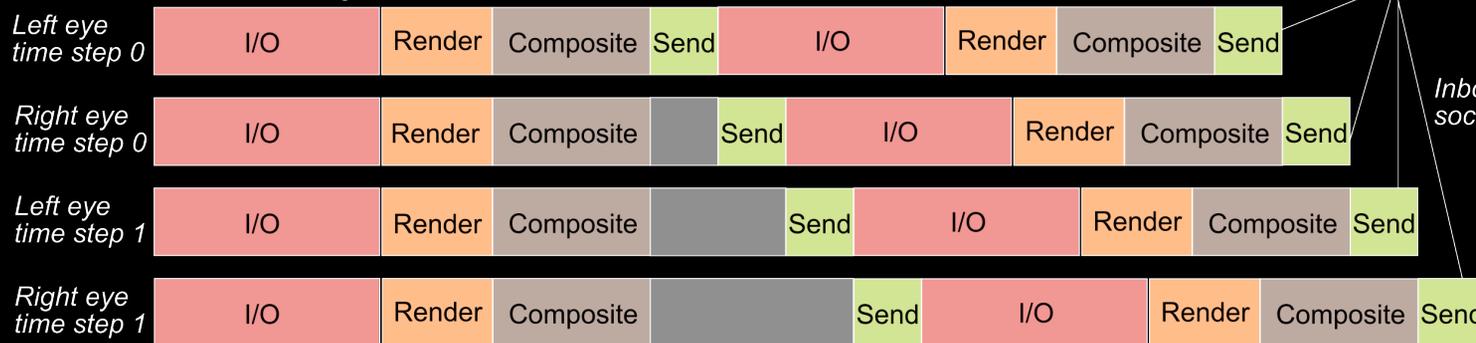
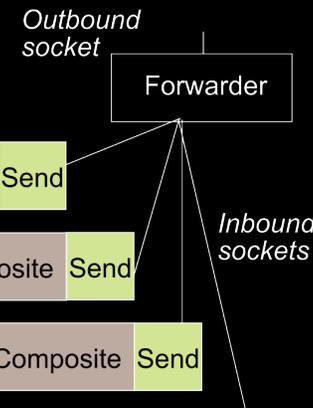
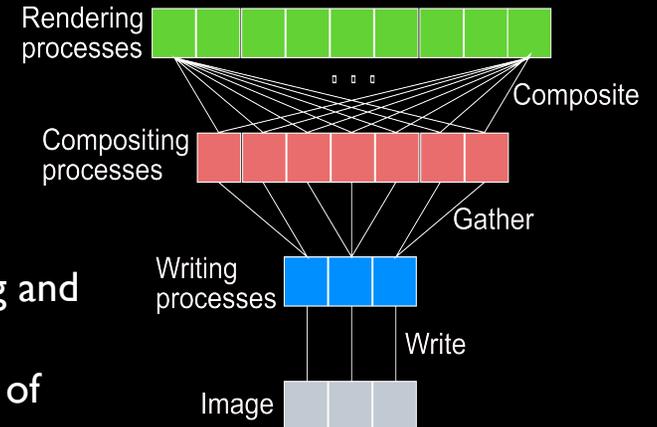
Other Optimizations

Parallel Pipelining and I/O Subsetting

Output Time for Varying Numbers of Writing Processes



Parallel image writing and I/O subsetting: Controlling number of total processes that perform parallel writing can boost performance.

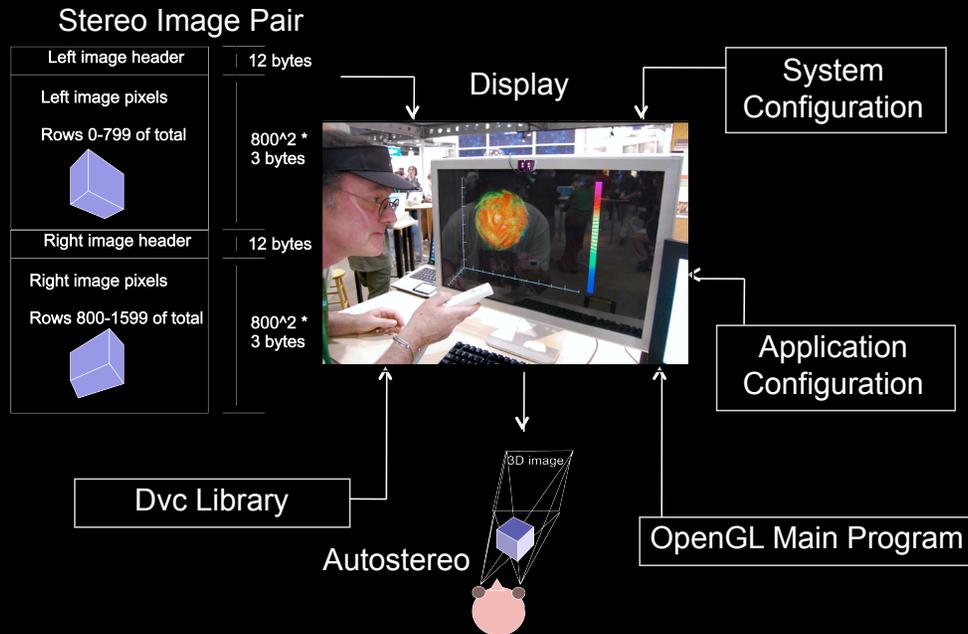


Parallel pipelining: I/O latency in a time series can be masked by visualizing multiple time steps in parallel pipelines. Each of the pipelines below is further parallelized among multiple nodes.

Other Optimizations

Stereo Volume Rendering

Display of Large-Scale Scientific Visualization. SPIE'09

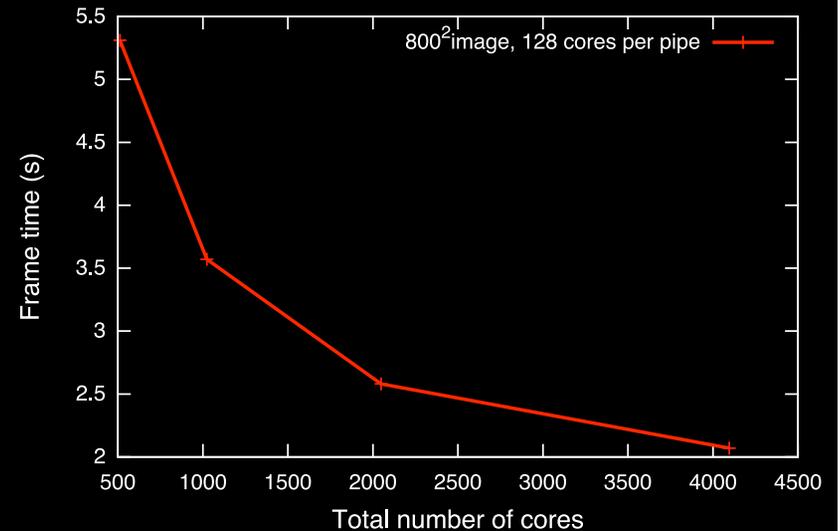


Stereo parallel volume rendering: The server (BG/P) computes stereo pairs of volume-rendered images and streams them to the client, which runs the dvc library to display them remotely in autostereo. End-to-end frame times of 2 s. per frame were achieved over a 3-hour demo from Argonne to Austin, TX.



Display devices and interaction techniques bring virtual environments to scientific visualization.

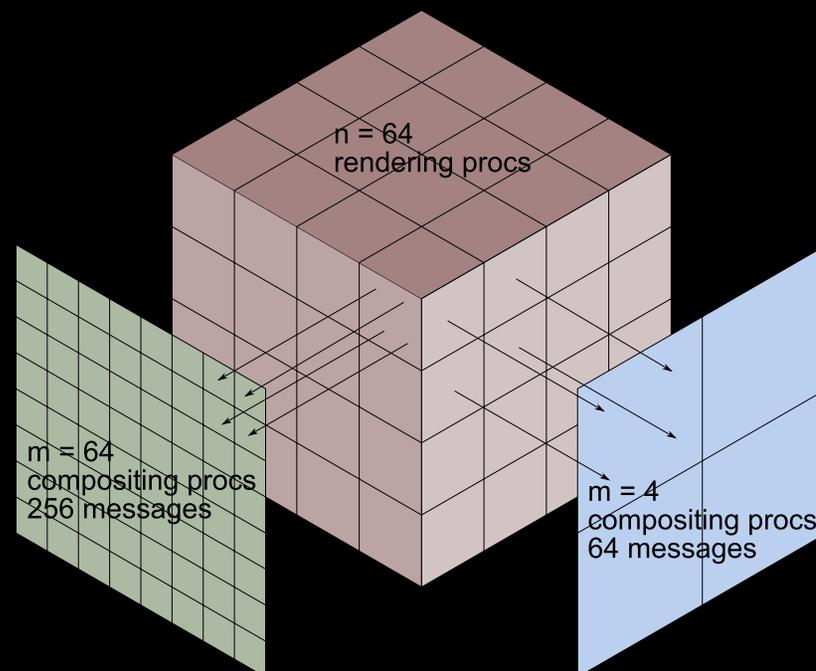
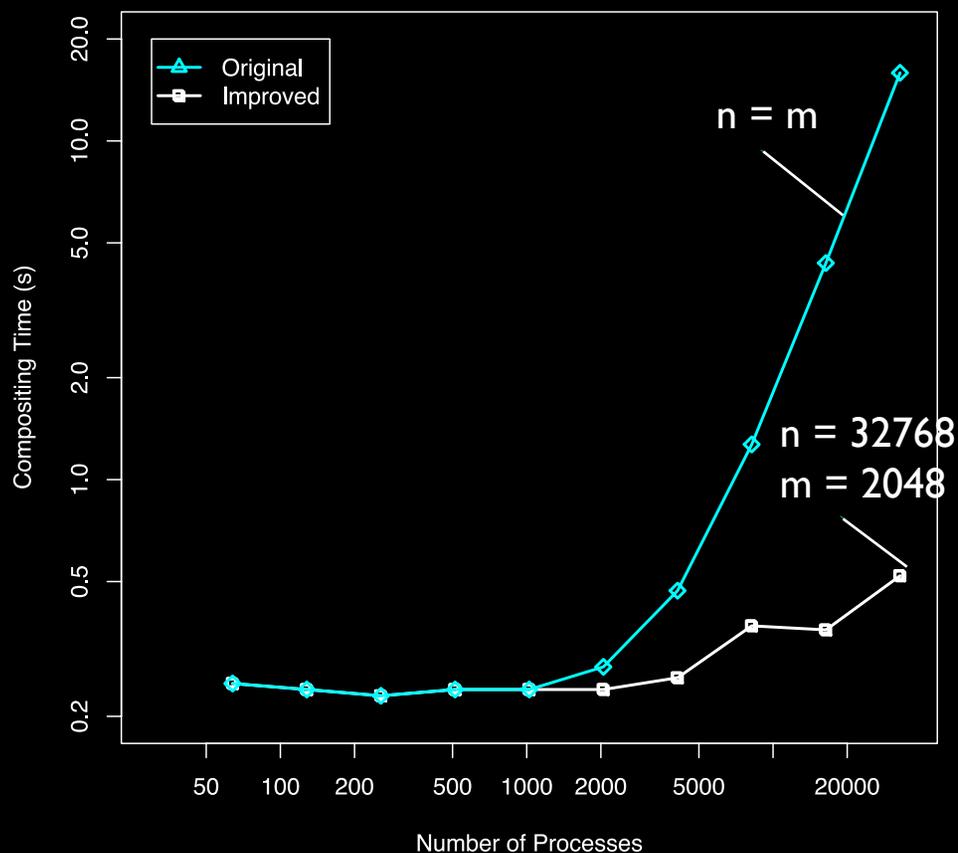
Stereo Performance, 864³ data



Other Optimizations

Reducing the number of compositing processes improves direct-send performance.

Compositing Time



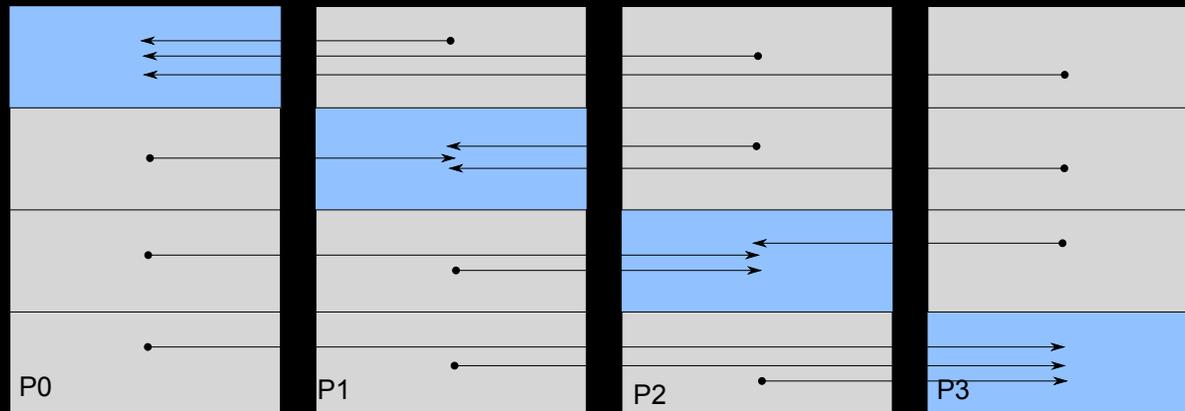
Usually in direct-send, $n = m$, but setting $m < n$ can reduce contention when n is large. On average, $O(m * n^{1/3})$ total messages, can get down to $O(n)$ if $m = n^{2/3}$.

Direct-send compositing time improved up to 30X. | 120^3 data volume, 1600^2 image size.

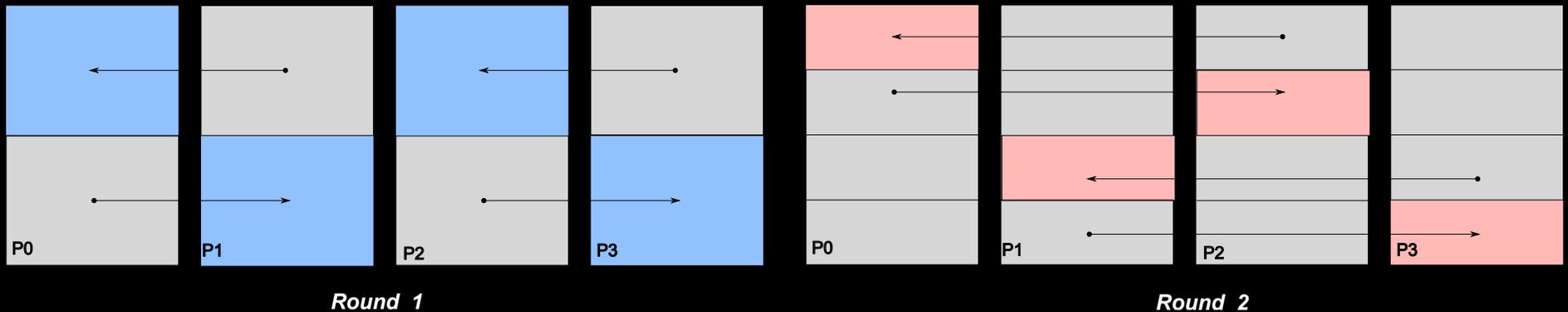
End-to-End Study of Parallel Volume Rendering on the IBM Blue Gene/P. ICPP'09

Image Compositing

Direct-send and binary swap



Direct-send: maximum parallelism but high number of small messages results in network contention

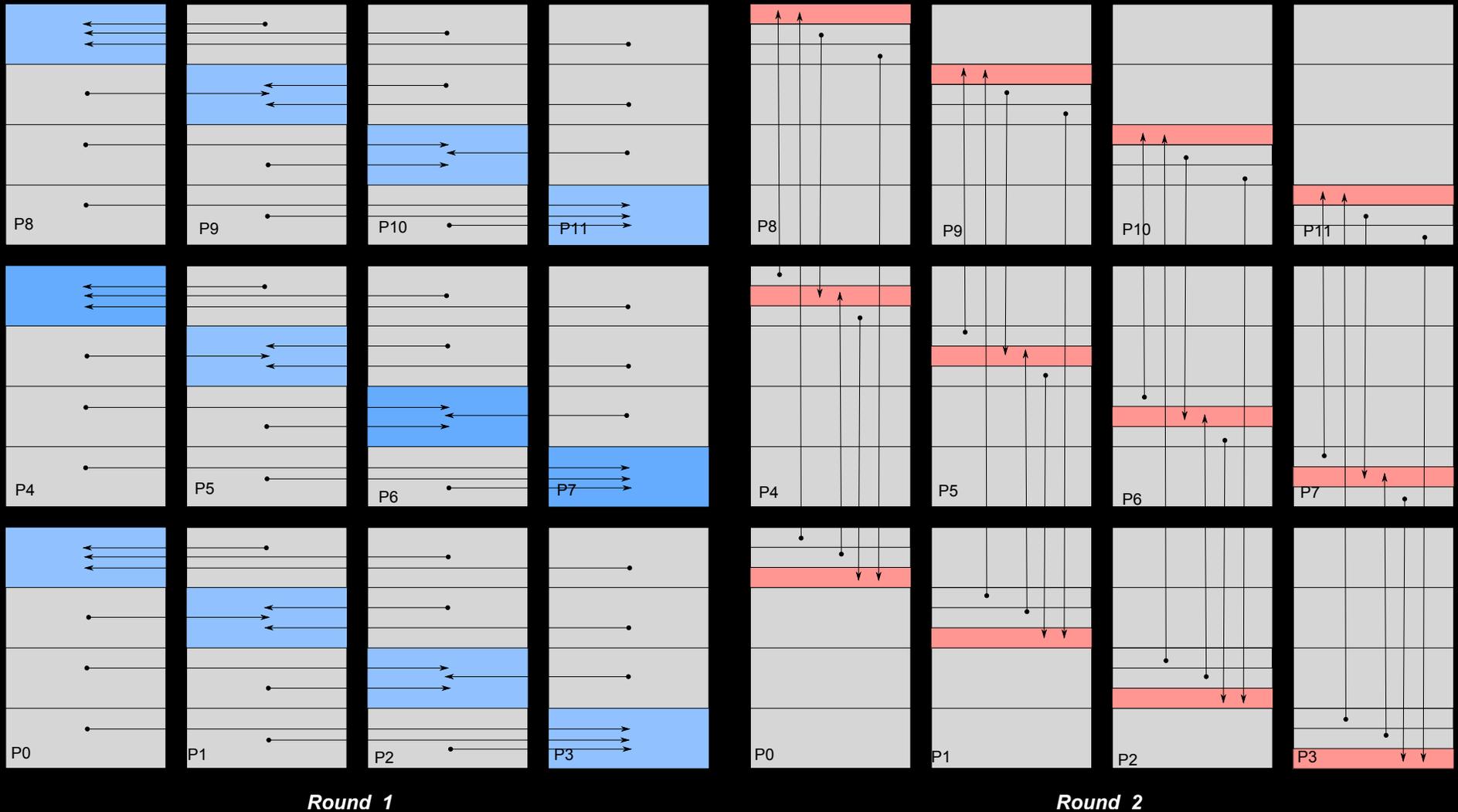


Binary swap: fewer messages over $\log_2 p$ rounds, $p =$ number of processes, power of 2, strictly synchronous

Radix-k Compositing

The best of both worlds

A Configurable Algorithm for Parallel Image
-Compositing Applications. SC09.

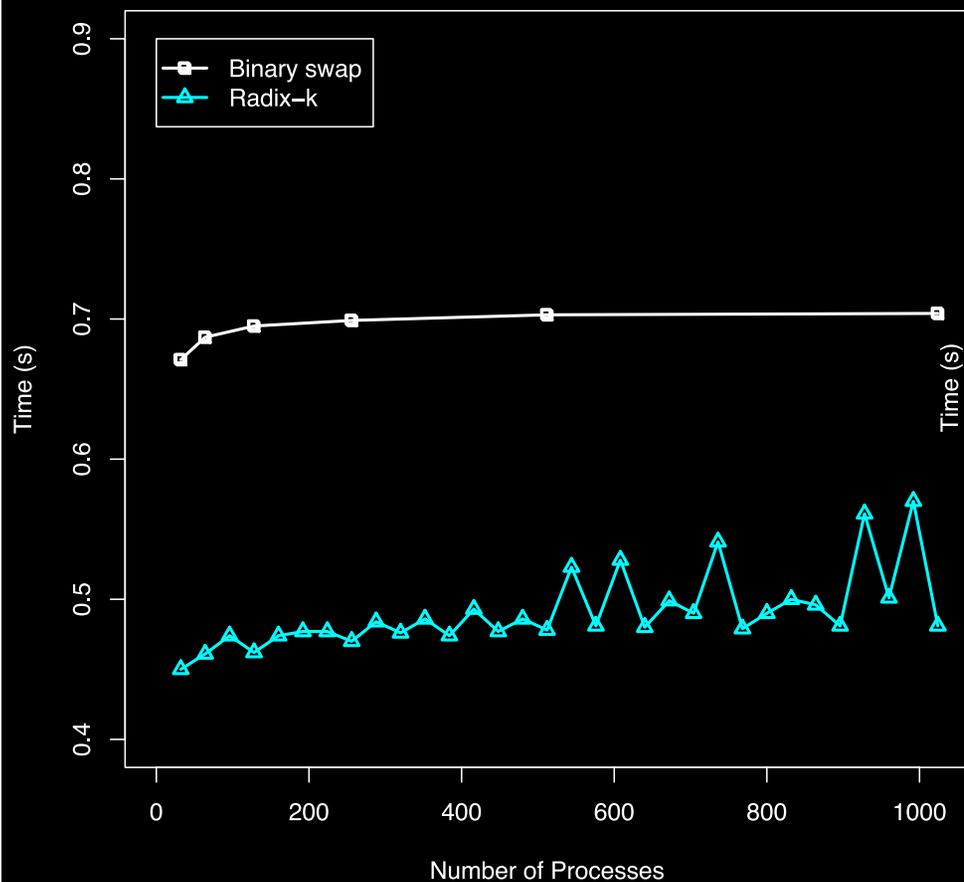


Radix-k: More parallel, less synchronous, managed contention, p does not need to be power of 2

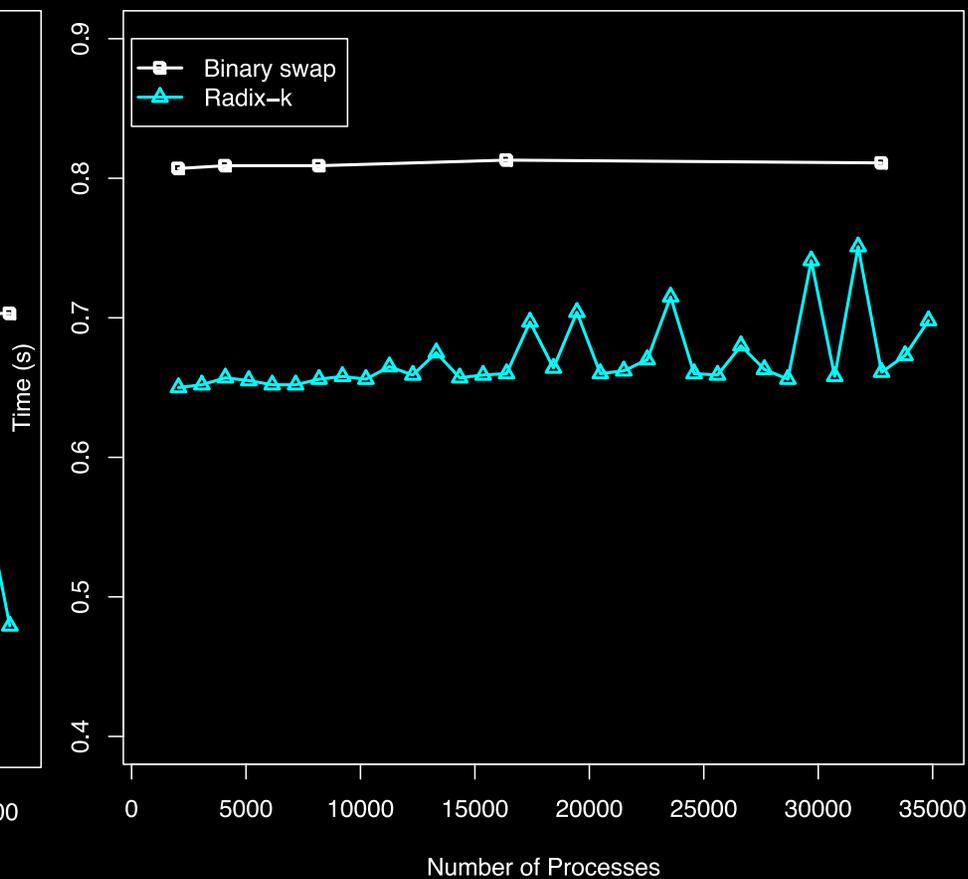
Radix-k Performance

From 32 to 35,000 processes on Blue Gene/P Intrepid

Compositing Time for 8 Mpx Image



Compositing Time for 8 Mpx Image



Tested at 1, 2, 4, and 8 Mpix. 1 pixel = 4 floats (16 bytes per pixel)
40% improvement over binary swap at a variety of process counts. Left: p varies from 32 to 1024 in steps of 32. Right: p continues from 1024 to 35,000 in steps of 1024.

Recap

Lessons learned and the road ahead

Successes

- Demonstrated scaling on large data and images
- Improved compositing
- Improved and benchmarked I/O

Ongoing

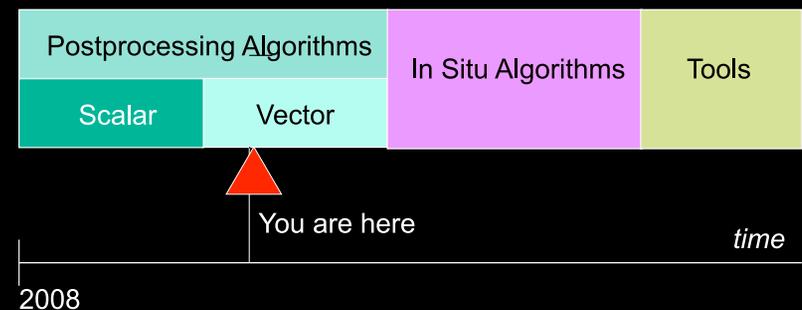
- Other algorithms and grid topologies
- In situ
- Adoption into tools and libraries

Take-away

- **HPC has appropriate resources for visualization:** massive parallelism, storage, and interconnect capability.

- **Visualization algorithms can be developed that scale** with the machine and problem size.

- **Embrace parallelism;** it is here to stay.



Further Reading

References

Peterka, T., Goodell, D., Ross, R., Shen, H.-W., Thakur, R.: A Configurable Algorithm for Parallel Image-Compositing Applications. Proceedings of SC09, Portland OR, November 2009.

Peterka, T., Yu, Hongfeng, Ross, R., Ma, K.-L., Latham, R.: End-to-End Study of Parallel Volume Rendering on the IBM Blue Gene/P. Proceedings of ICPP'09, Vienna, Austria, September 2009.

Peterka, T., Ross, R. B., Shen, H.-W., Ma, K.-L., Kendall, W., Yu, H.: Parallel Visualization on Leadership Computing Resources. Journal of Physics: Conference Series SciDAC 2009, June 2009.

Peterka, T., Ross, R., Yu, H., Ma, K.-L., and Girado, Javier: Autostereoscopic Display of Large-Scale Scientific Visualization. Proceedings of IS&T / SPIE SD&A XX Conference, San Jose CA, January 2009.

Peterka, T., Ross, R., Yu, H., Ma, K.-L.: Assessing Improvements to the Parallel Volume Rendering Pipeline at Large Scale. SC08 Ultrascale Visualization Workshop, Austin TX, November 2008.

Ross, R. B., Peterka, T., Shen, H.-W., Hong, Y., Ma, K.-L., Yu, H., Moreland, K.: Parallel I/O and Visualization at Extreme Scale. Journal of Physics: Conference Series SciDAC 2008, July 2008.

Peterka, T., Yu, H., Ross, R., Ma, K.-L.: Parallel Volume Rendering on the IBM Blue Gene/P. Proceedings of Eurographics Symposium on Parallel Graphics and Visualization 2008 (EGPGV'08) Crete, Greece, April 2008.



Argonne
NATIONAL
LABORATORY

... for a brighter future



SciDAC Institute for Ultrascale Visualization

www.ultravis.org



U.S. Department
of Energy

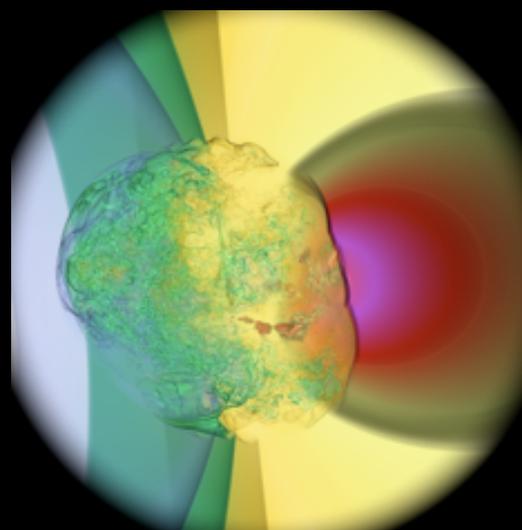
UChicago ►
Argonne_{LLC}



U.S. DEPARTMENT OF ENERGY

A U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC

Scalable Approaches to Analysis of Scientific Data



Acknowledgments:

Hongfeng Yu, Wes Kendall, Rob Latham,
Dave Goodell, Kwan-Liu Ma, Rob Ross,
Han-Wei Shen, Rajeev Thakur
John Blondin, Tony Mezzacappa
Argonne and Oak Ridge Leadership
Computing Facilities
US DOE SciDAC UltraVis Institute

Tom Peterka

tpeterka@mcs.anl.gov

Mathematics and Computer Science Division