TACC
TEXAS ADVANCED COMPUTING CENTER
WWW.TACC.UTEXAS.EDU

TEXAS
The University of Texas at Austin

Visualization Communication at Scale: Opportunities for Improved Efficiency

6th Annual MVAPICH User Group (MUG) Meeting

August 7th, 2018

PRESENTED BY:

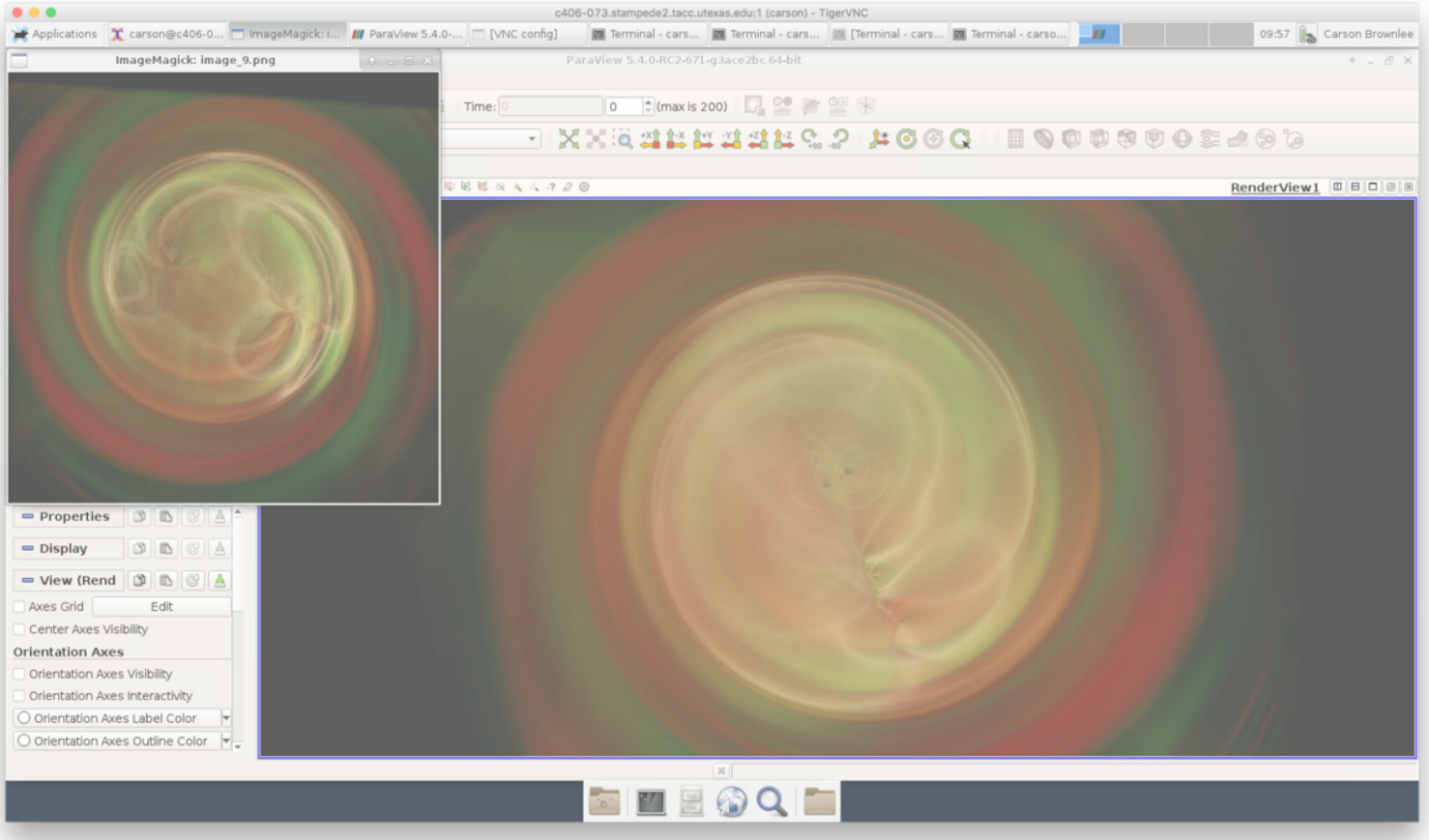
Paul A. Navrátil, Ph.D.

Director of Visualization

pnav@tacc.utexas.edu

Presentation Outline

- In Situ Overview
- Opportunities for MPI
- Performance Analysis Study for Tiled Sort-Last Rendering
- Discussion



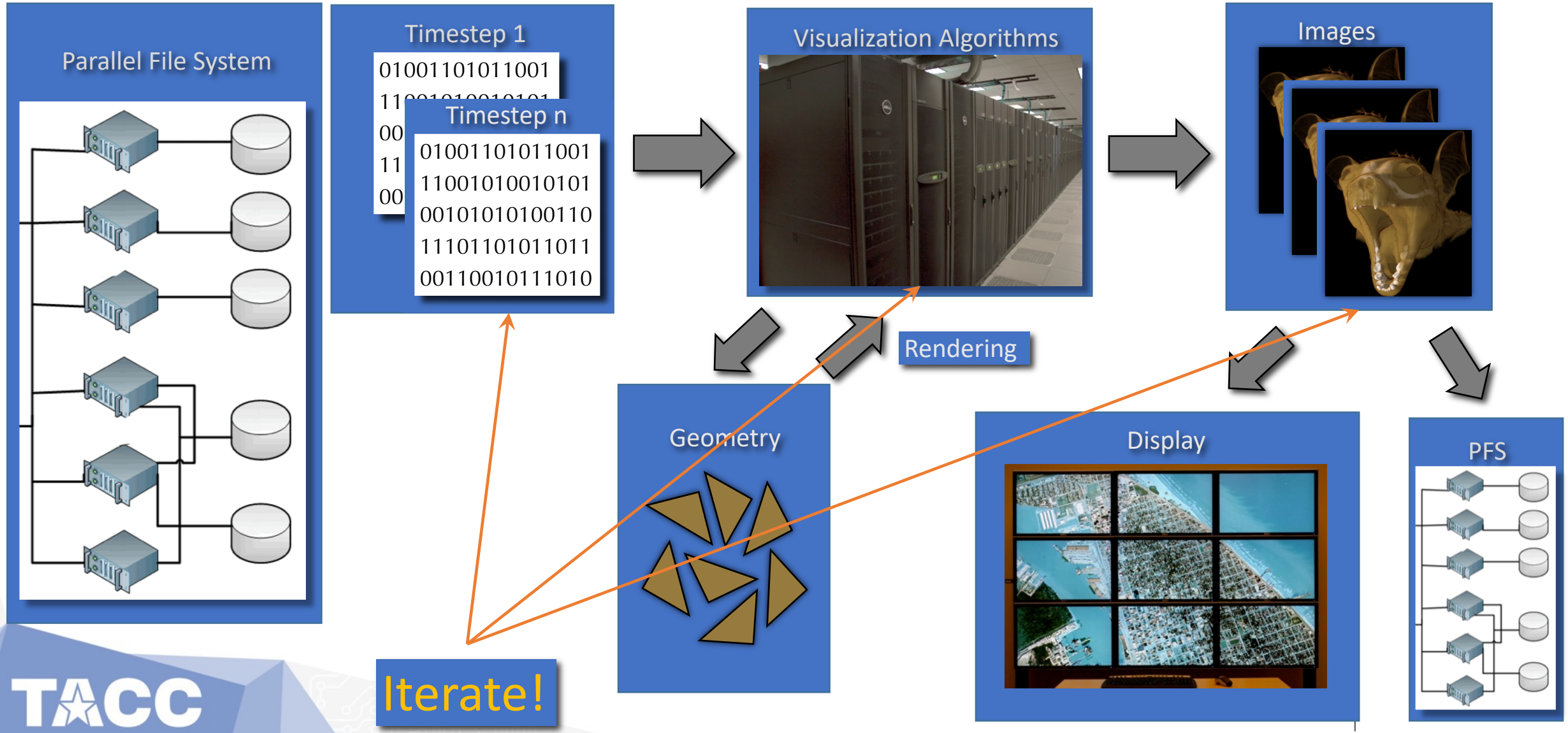
TACC

In Situ Visualization Overview

Why Software-Defined Visualization?

File Size	100 Gbps	10 Gbps	1 Gbps	300 Mbps	54 Mbps
1 GB	< 1 sec	1 sec	10 sec	35 sec	2.5 min
1 TB	~100 sec	~17 min	~3 hours	~10 hours	~43 hours
1 PB	~1 day	~12 days	~121 days	>1 year	~5 years

Typical Post-Hoc Visualization Workflow

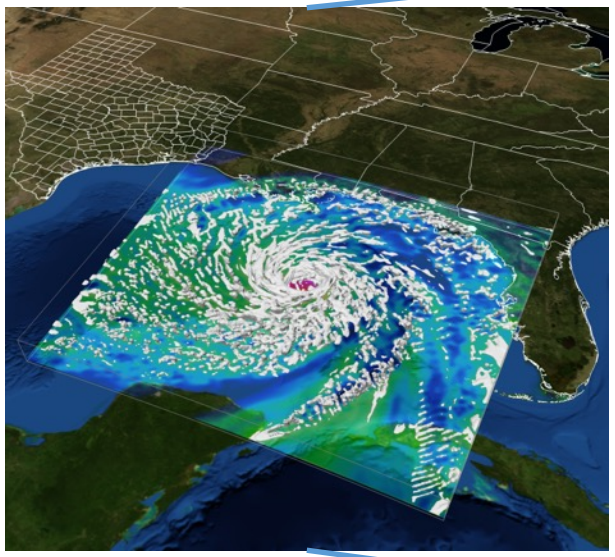


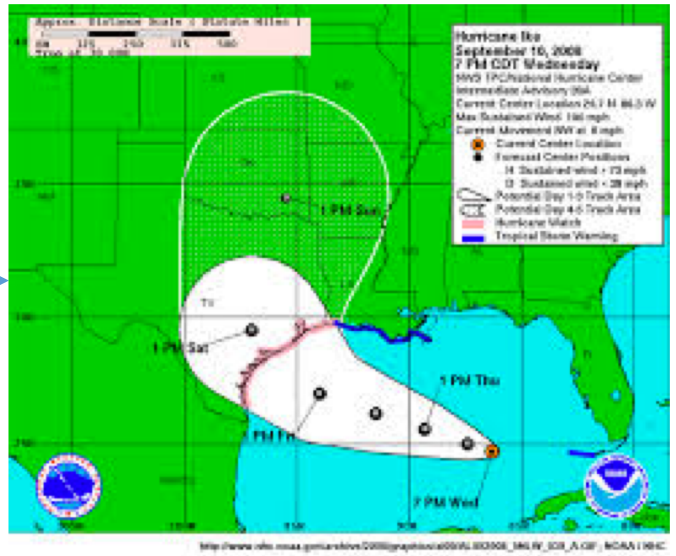
Why In Situ Visualization?

File Size	1000 GBps	100 GBps	10 GBps	1 GBps
1 TB	1 sec	~ 10 sec	~ 2 min	~ 17 min
1 PB	~ 17 min	~ 3 hours	~ 1 day	12 days
1 XB	12 days	124 days	3 ½ years	34 years

Why In Situ Visualization?

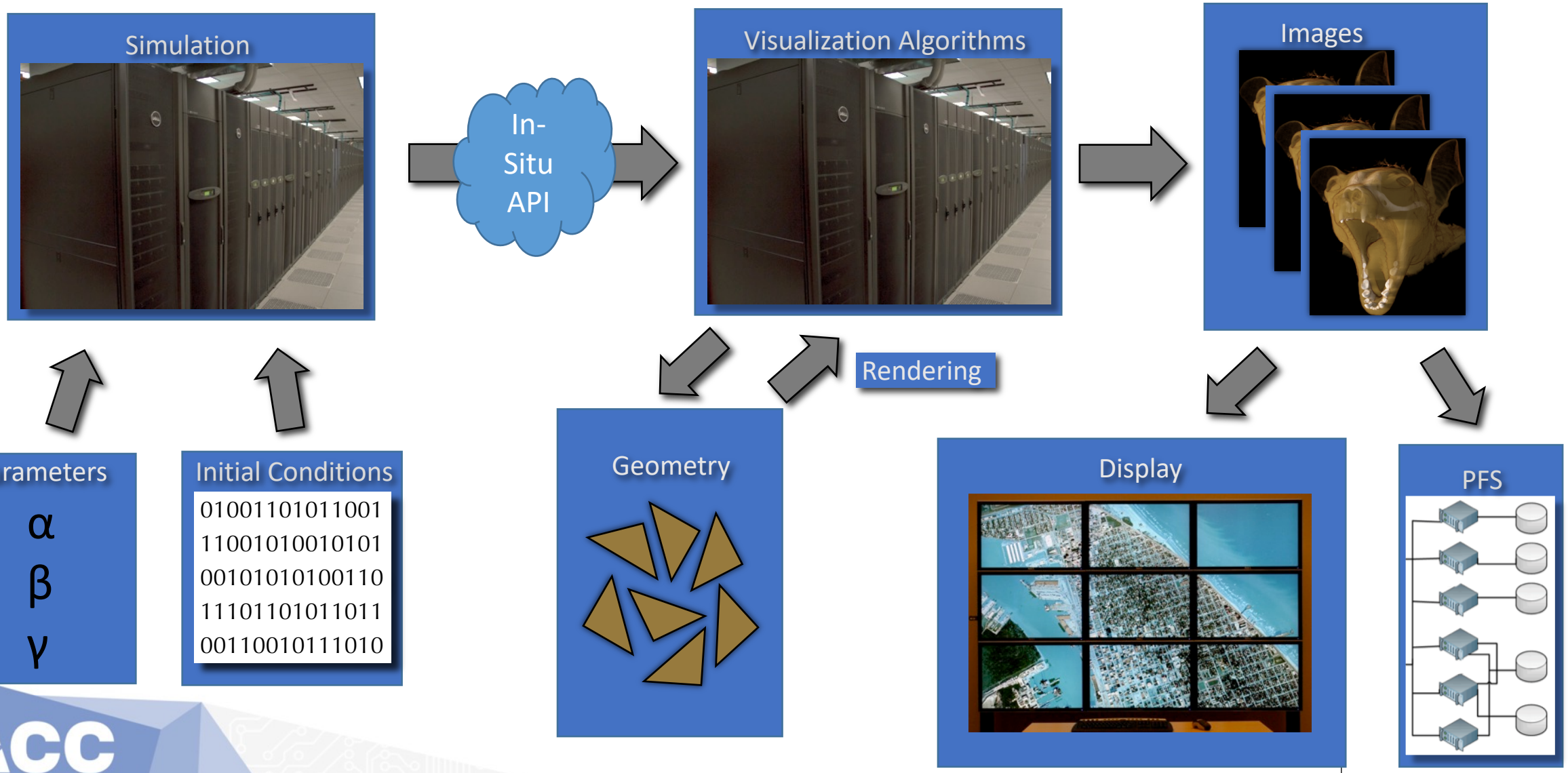
High-frequency writes





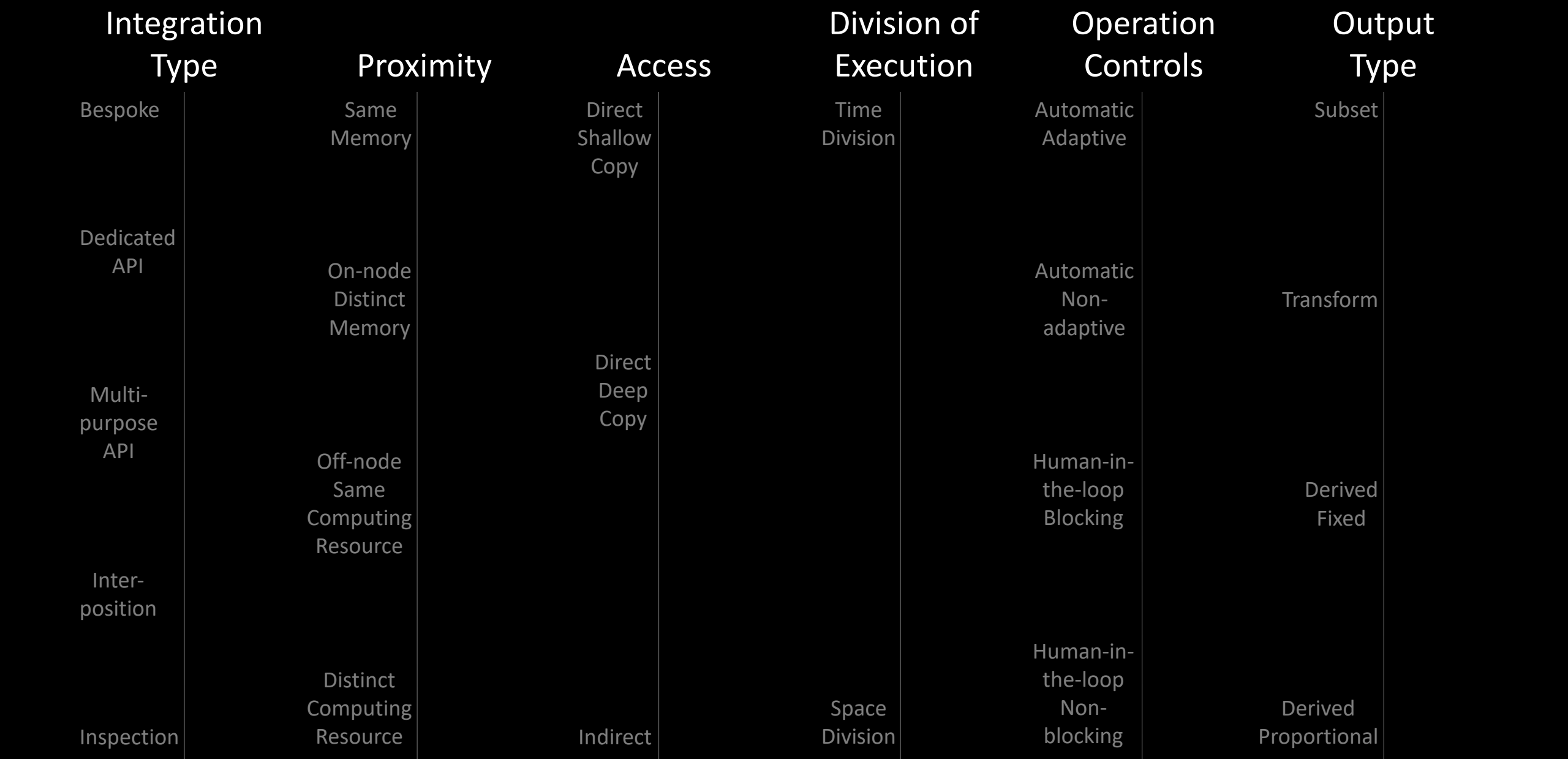
Low-frequency writes

In-Situ Visualization Workflow

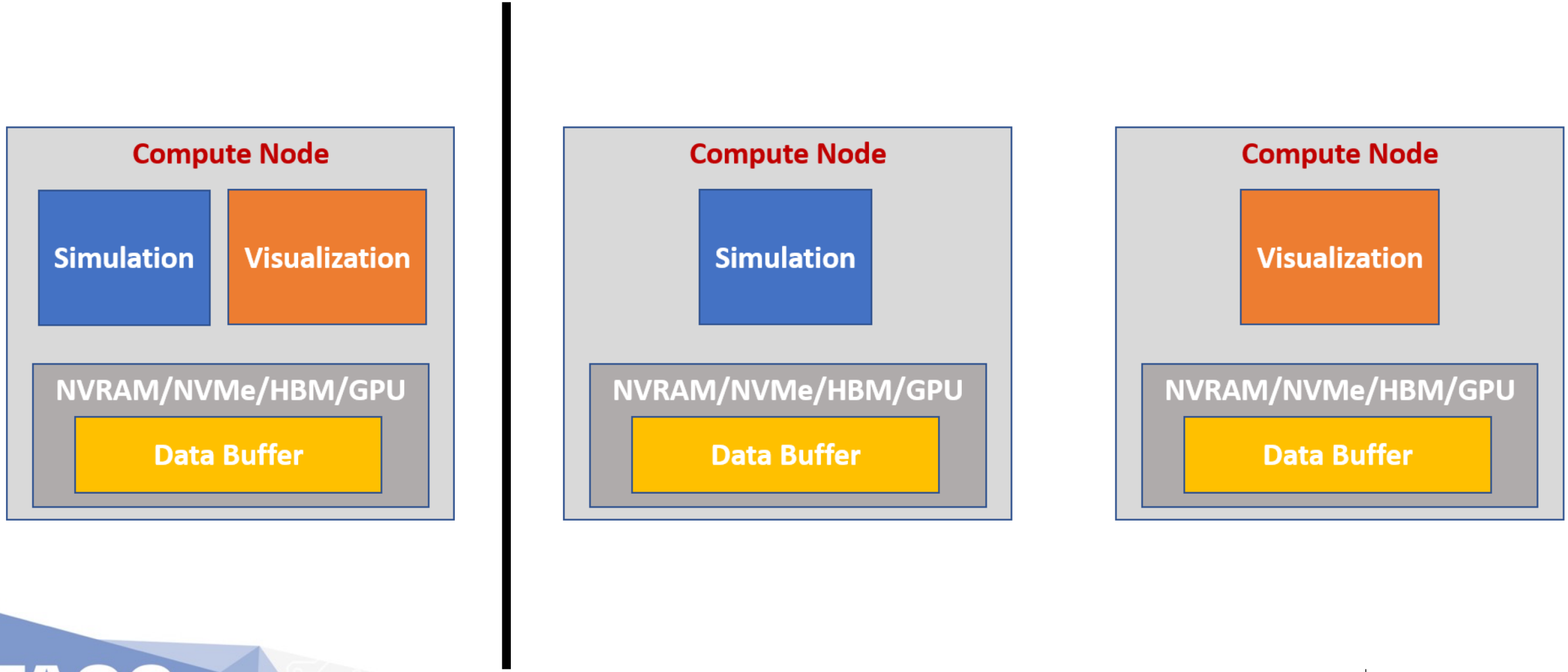


TACC

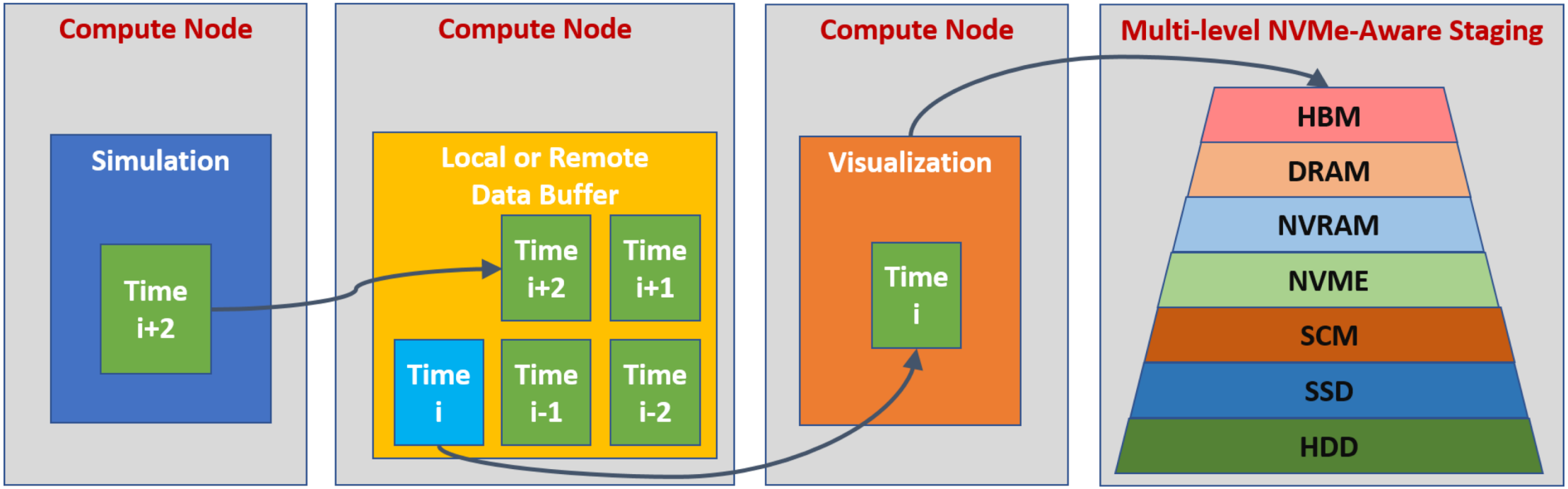
In Situ Terminology Project (courtesy Ken Moreland, Sandia)



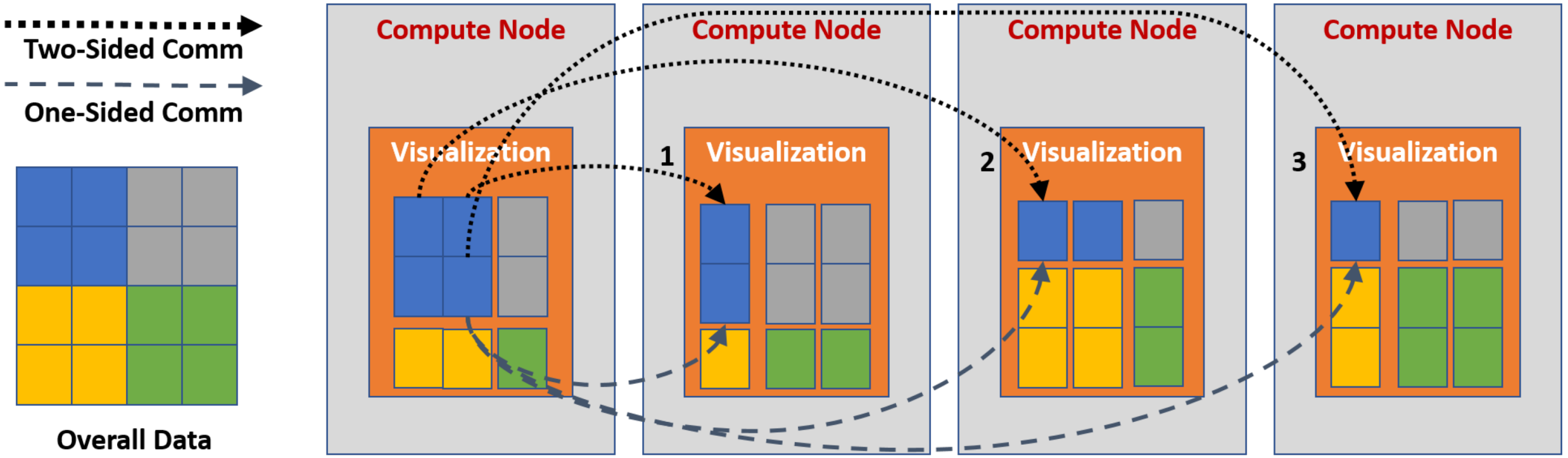
MPI Opportunity – Efficient Data Sharing



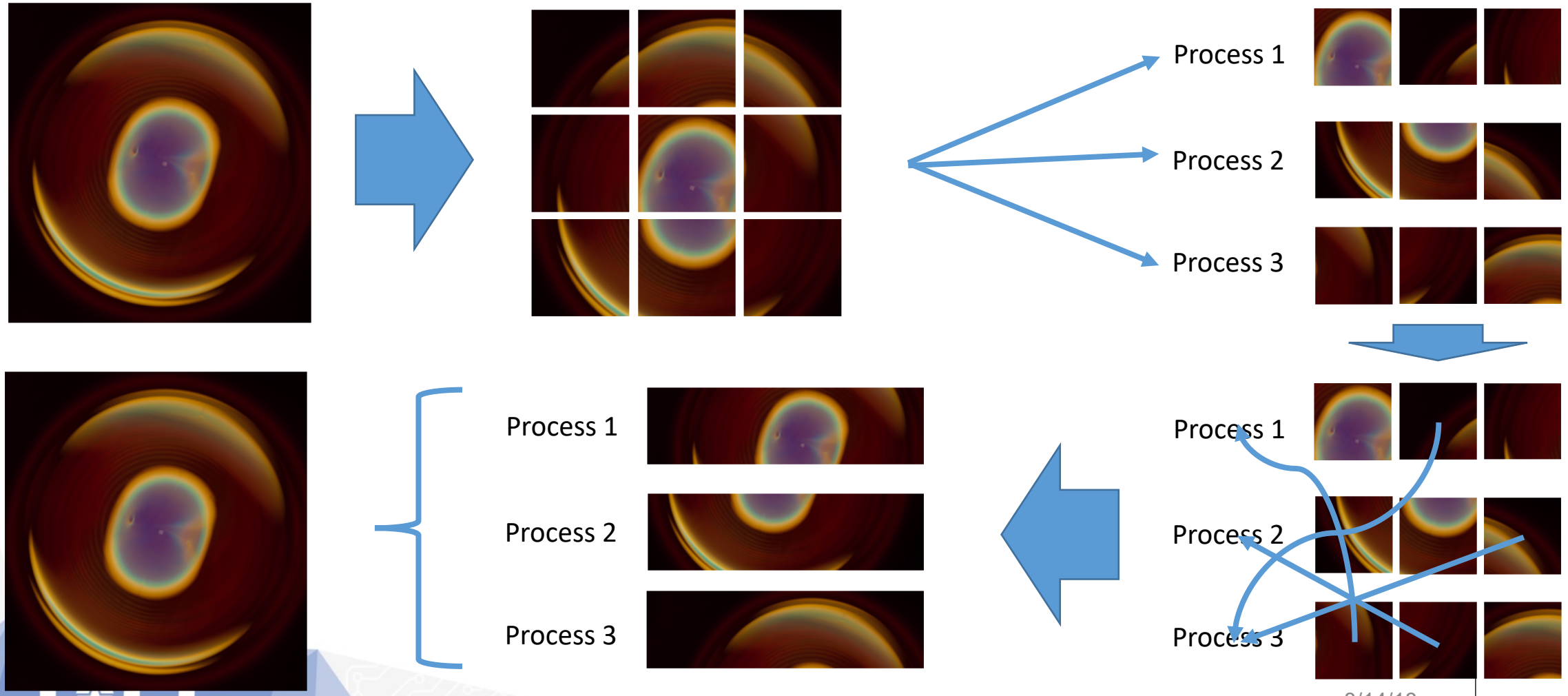
MPI Opportunity – Feature Evolution Analysis

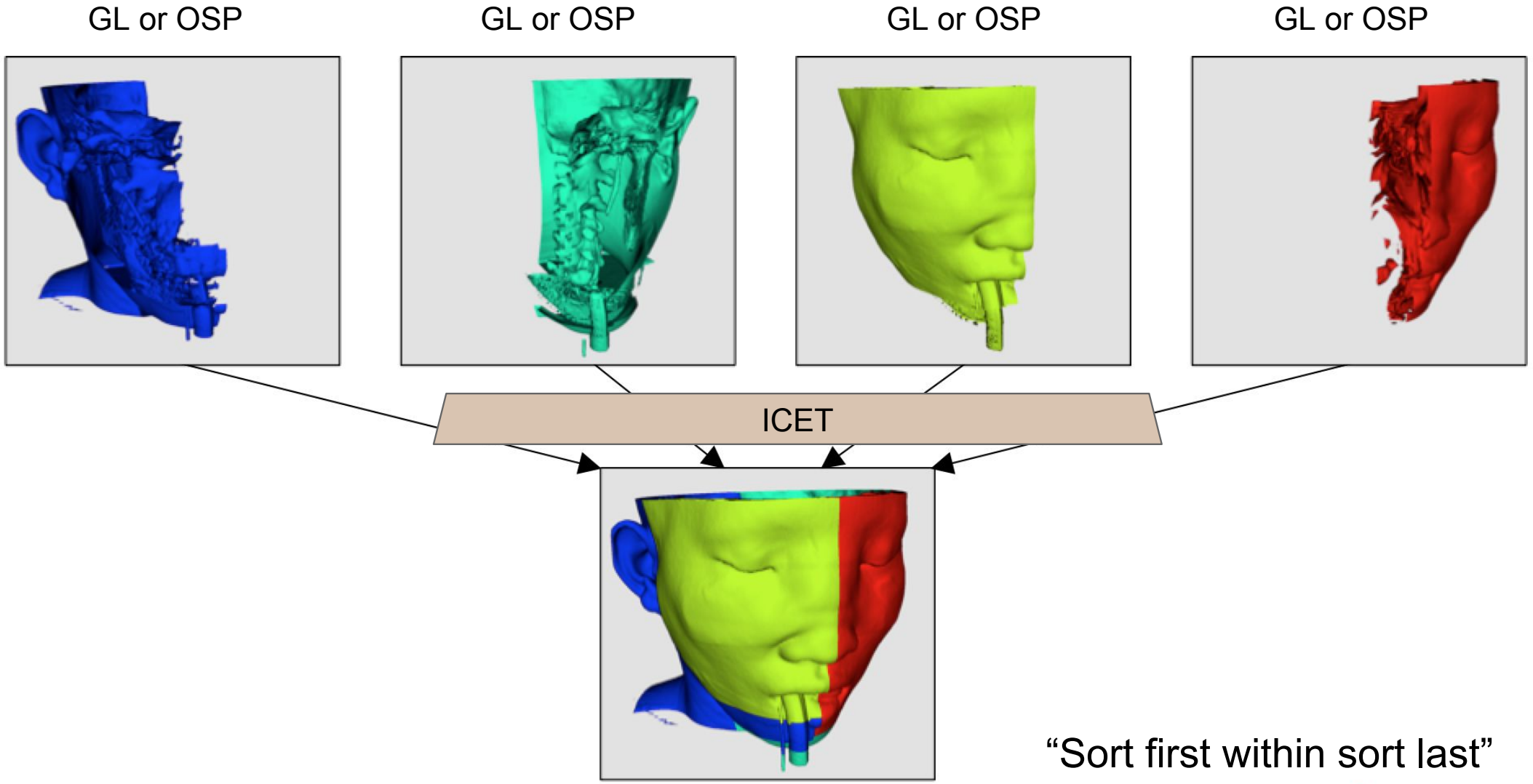


MPI Opportunity – Efficient Data Exchange

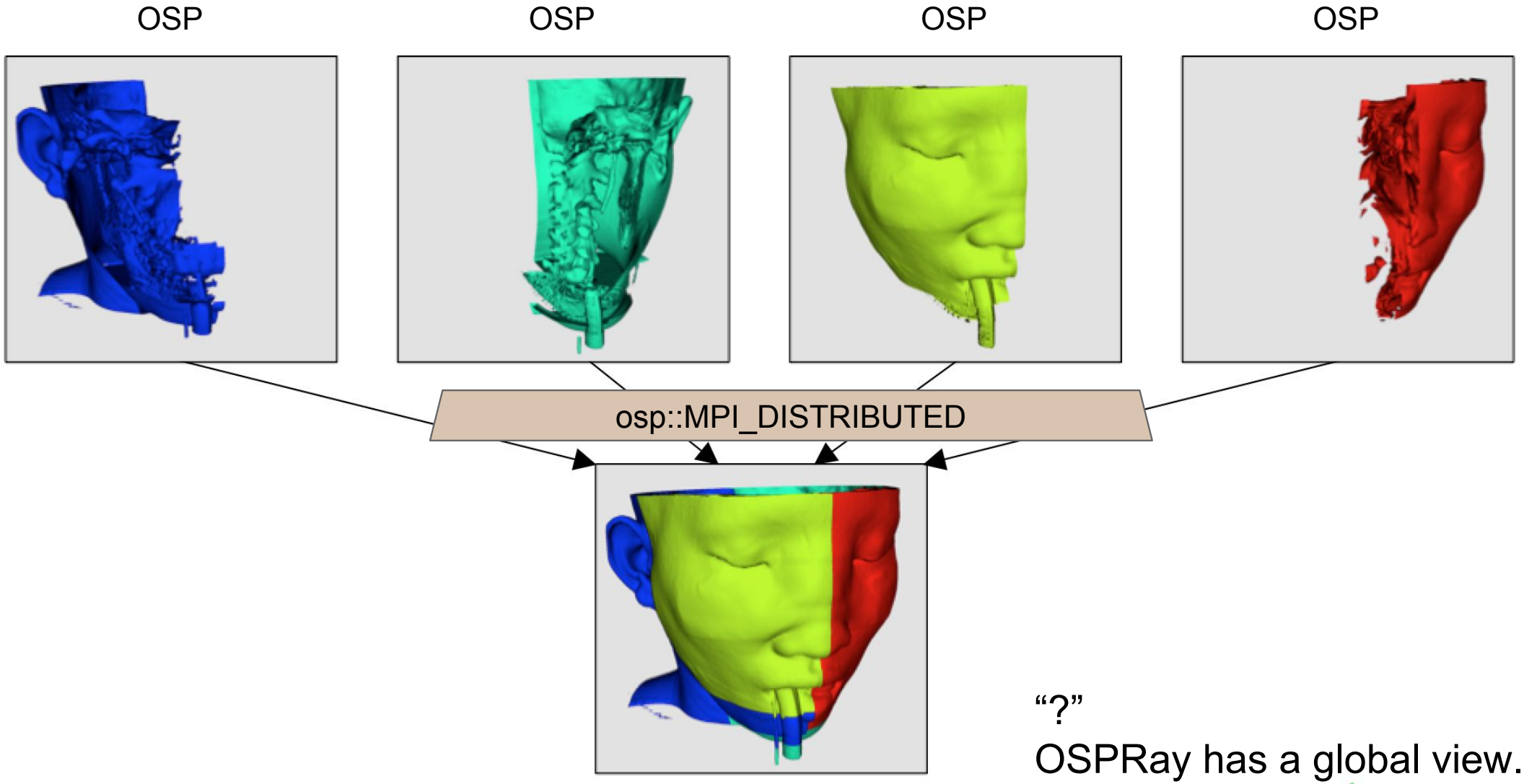


Performance Study– Tiled Sort-Last Rendering in Intel OSPRay (data courtesy Will Usher @ Intel)





“Sort first within sort last”



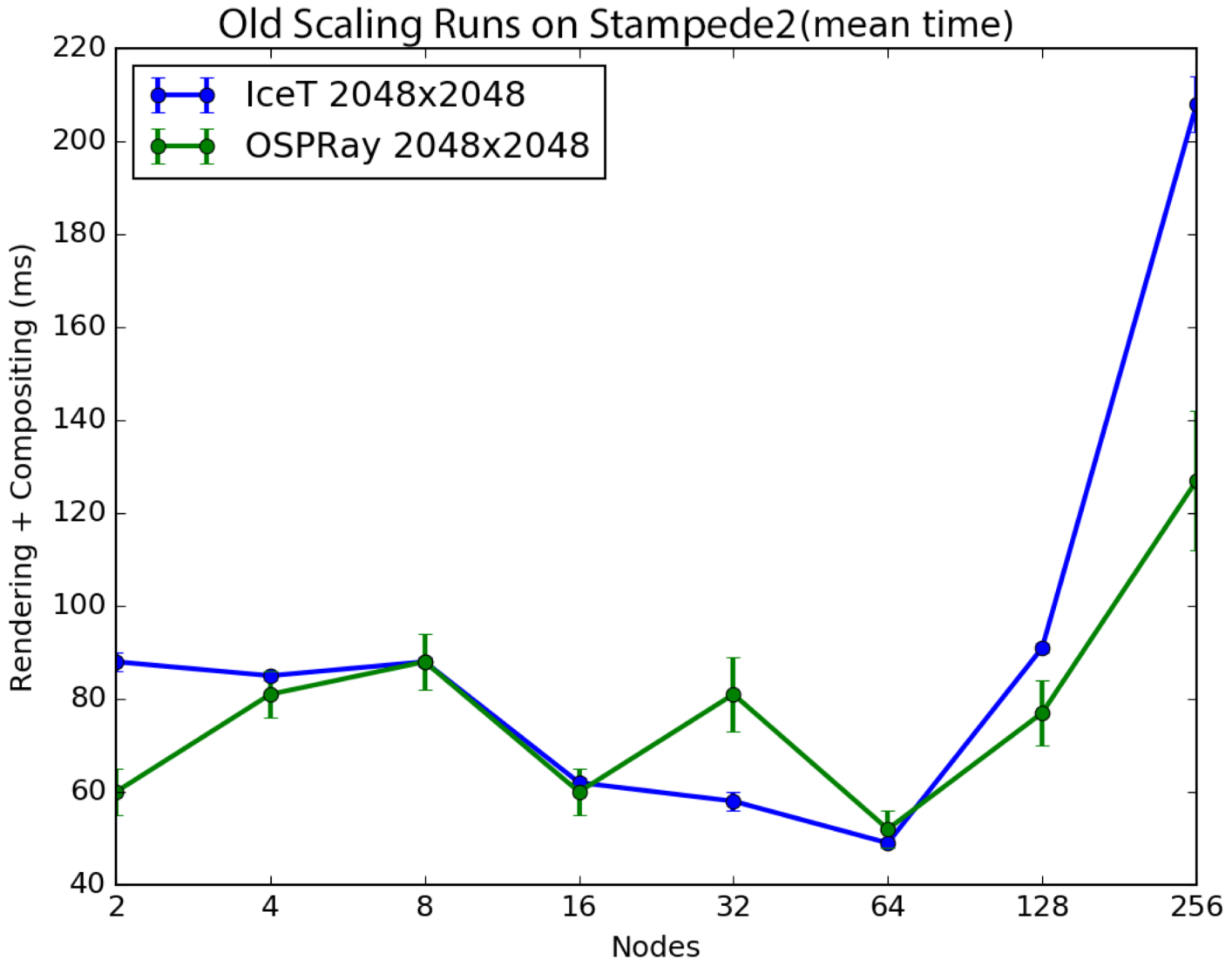
"?"

OSPRay has a global view.

Experiment Configuration

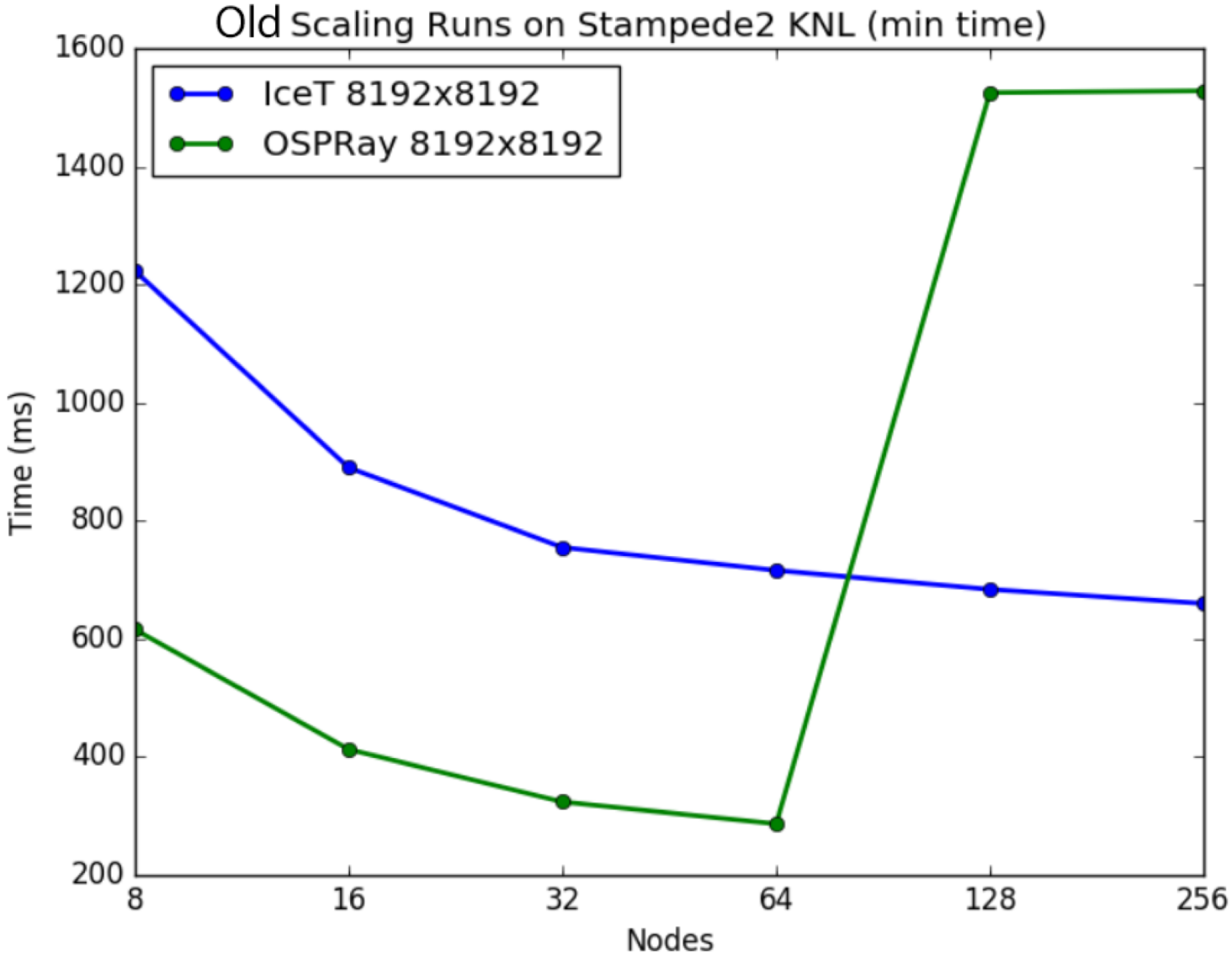
- Tested on TACC Stampede2 SKX / KNL, Argonne Theta
- Image sizes: 2048 x 2048, 8192 x 8192
- Each tile 64 x 64 pixels, ~ 81KB per payload (RGBA, depth)
- Tiles distributed point-to-point (MPI_Isend / MPI_Irecv)
- Partial tiles “gathered” point-to-point
- Fully rendered tiles sent to master point-to-point as completed
- Threads switch from rendering to compositing asynchronously
- Compare against IceT compositing library performance

Stampede2 KNL – 2K x 2K



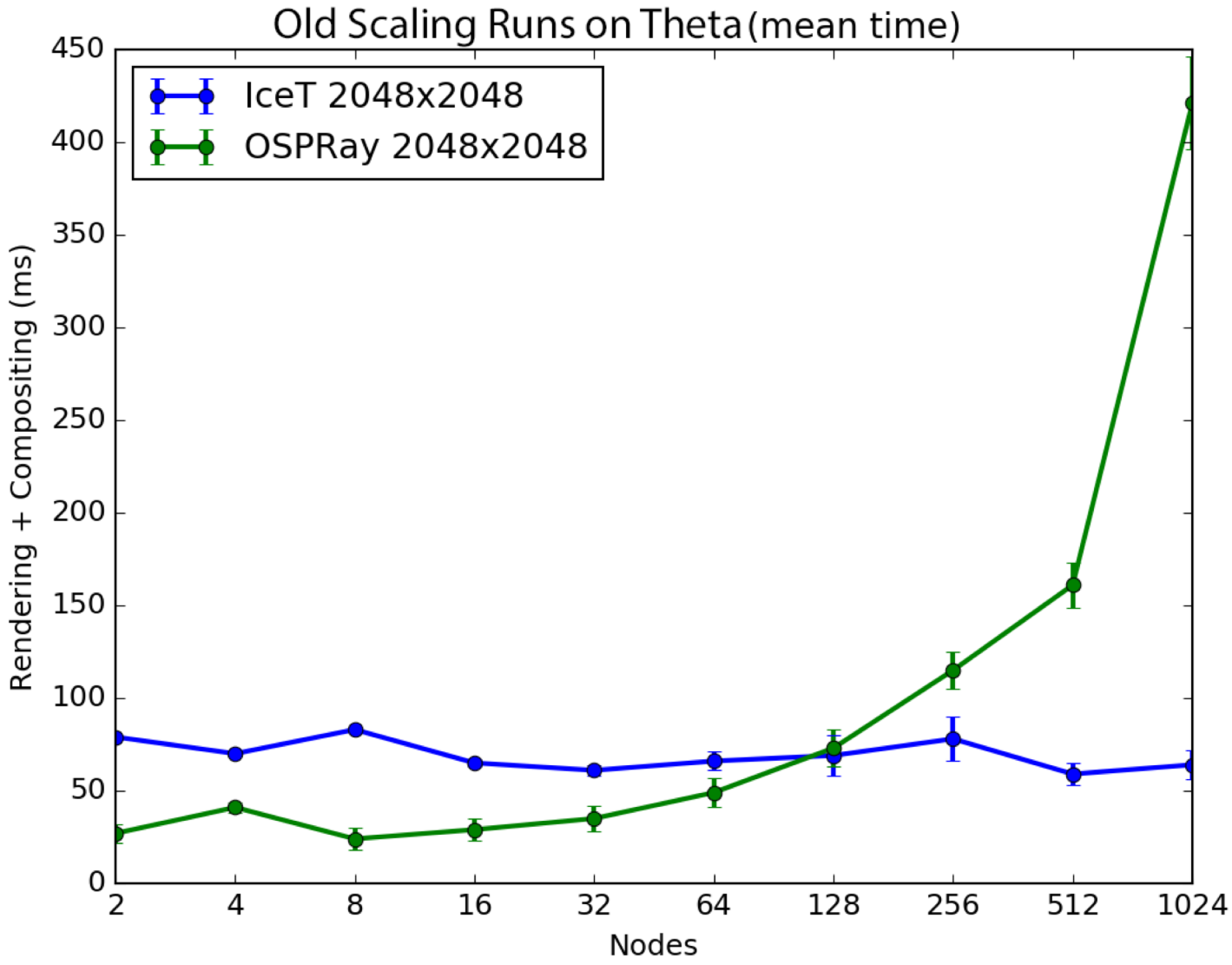
TACC

Stampede2 KNL – 8K x 8K

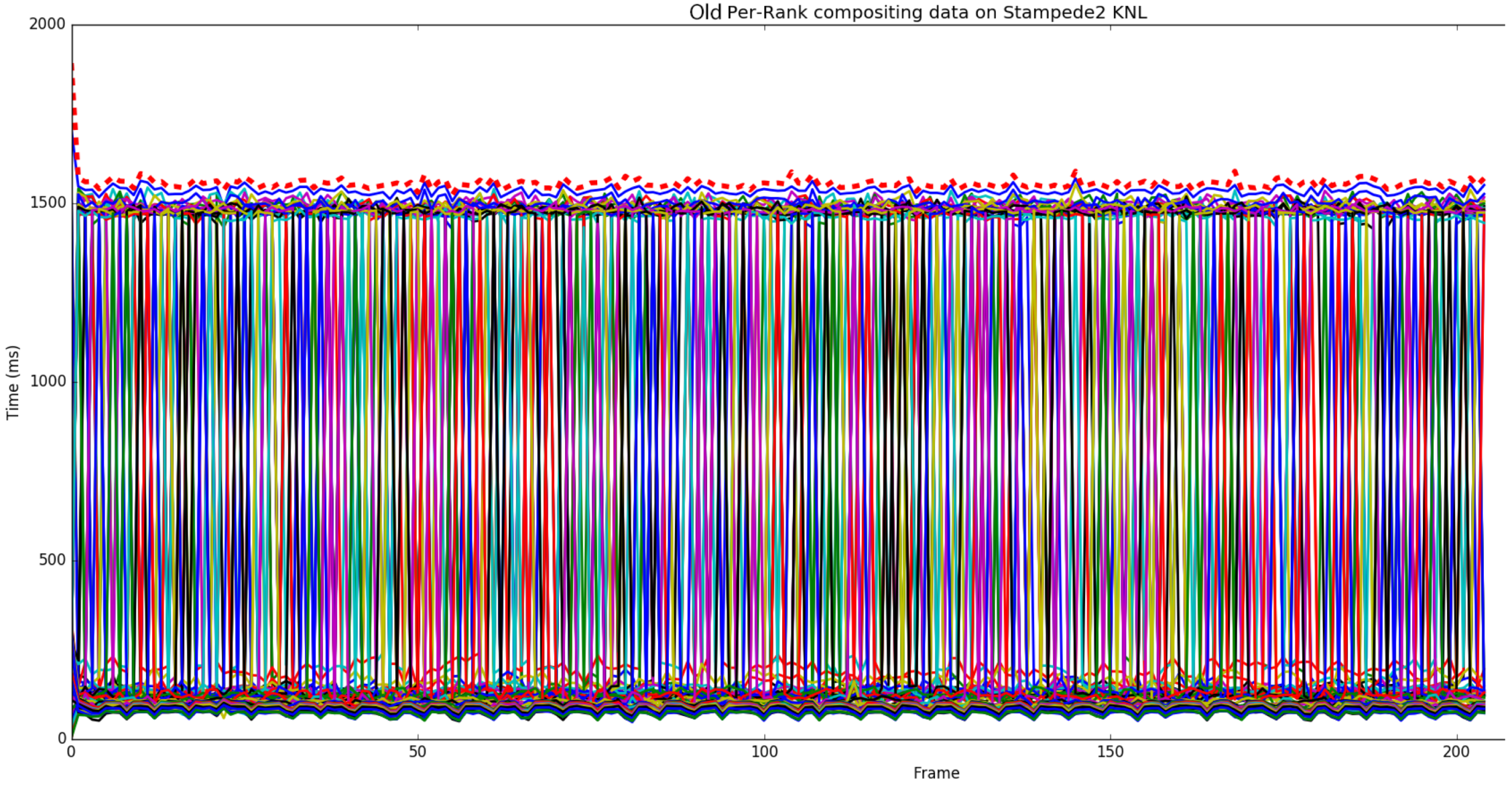


TACC

Theta KNL – 2K x 2K



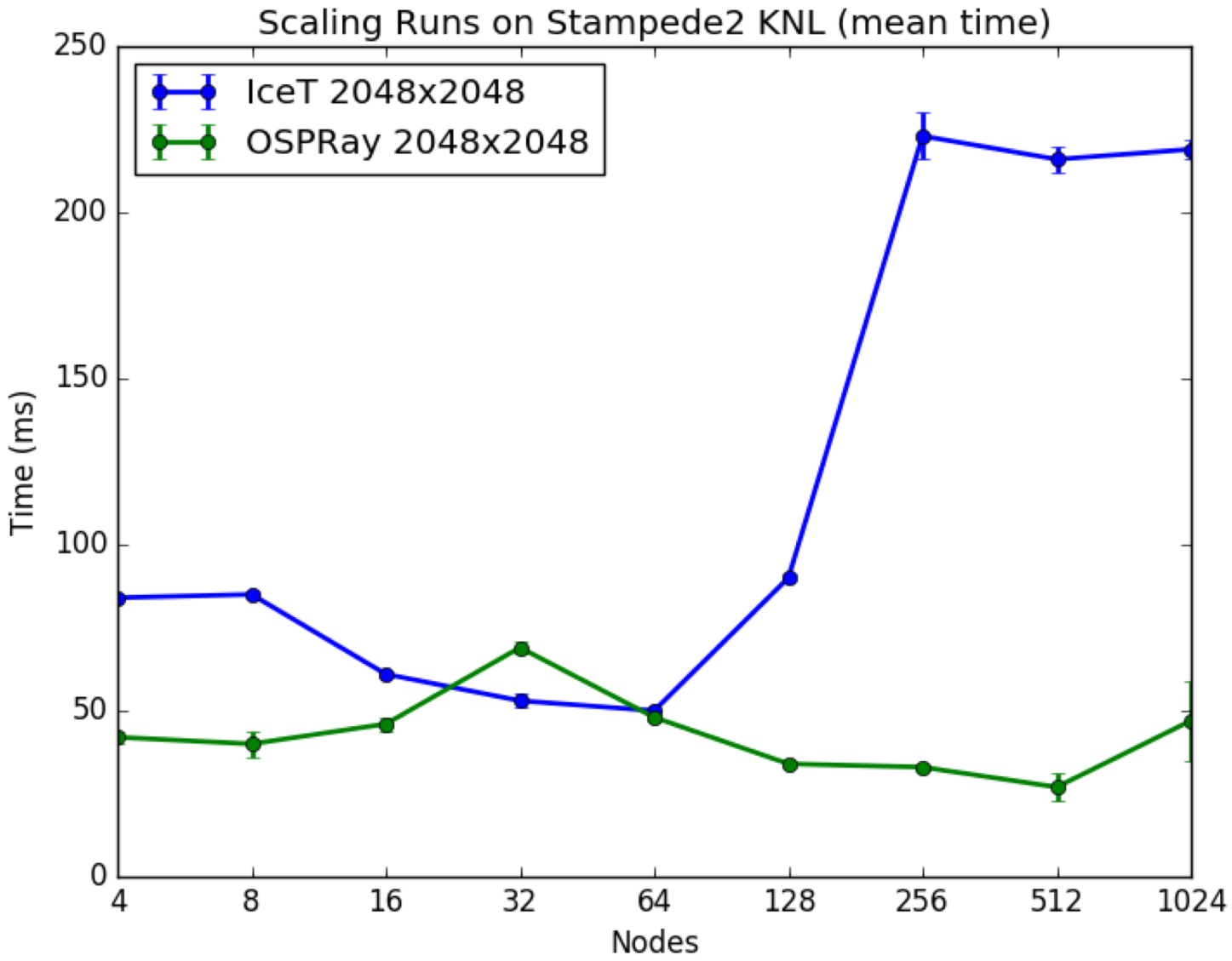
Stampede2 KNL – Compositing per rank
128 nodes @ 8K x 8K



So what's happening?

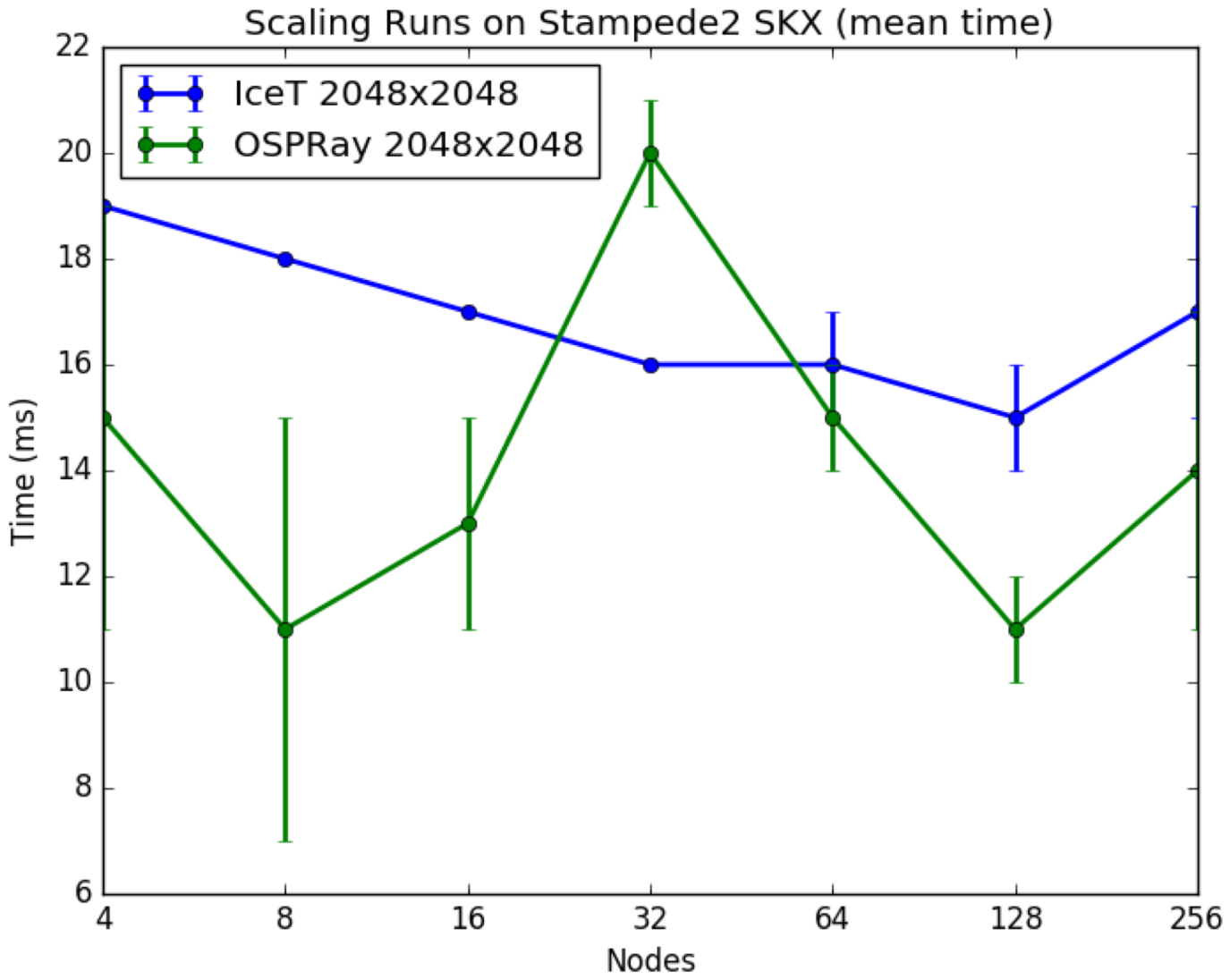
- Message size exceeds eager buffer threshold?
  - Increasing eager buffer size helps a bit at smaller image sizes, still see issues at 8K x 8K
- Message volume too high?
  - Maybe, though seems not (e.g. 8K @ 128 nodes = 128 tiles)
- Message size x message volume = ☹?
  - Perhaps we're on to something…
  - Compress messages with Google's Snappy algorithm
    - Per tile: 81.9 KB -> 2.4 KB
    - Image @ 8K: 268.6 MB -> 13.4 MB
  - Fully rendered tiles now sent to master via MPI_Allgatherv

Stampede2 KNL – 2K x 2K

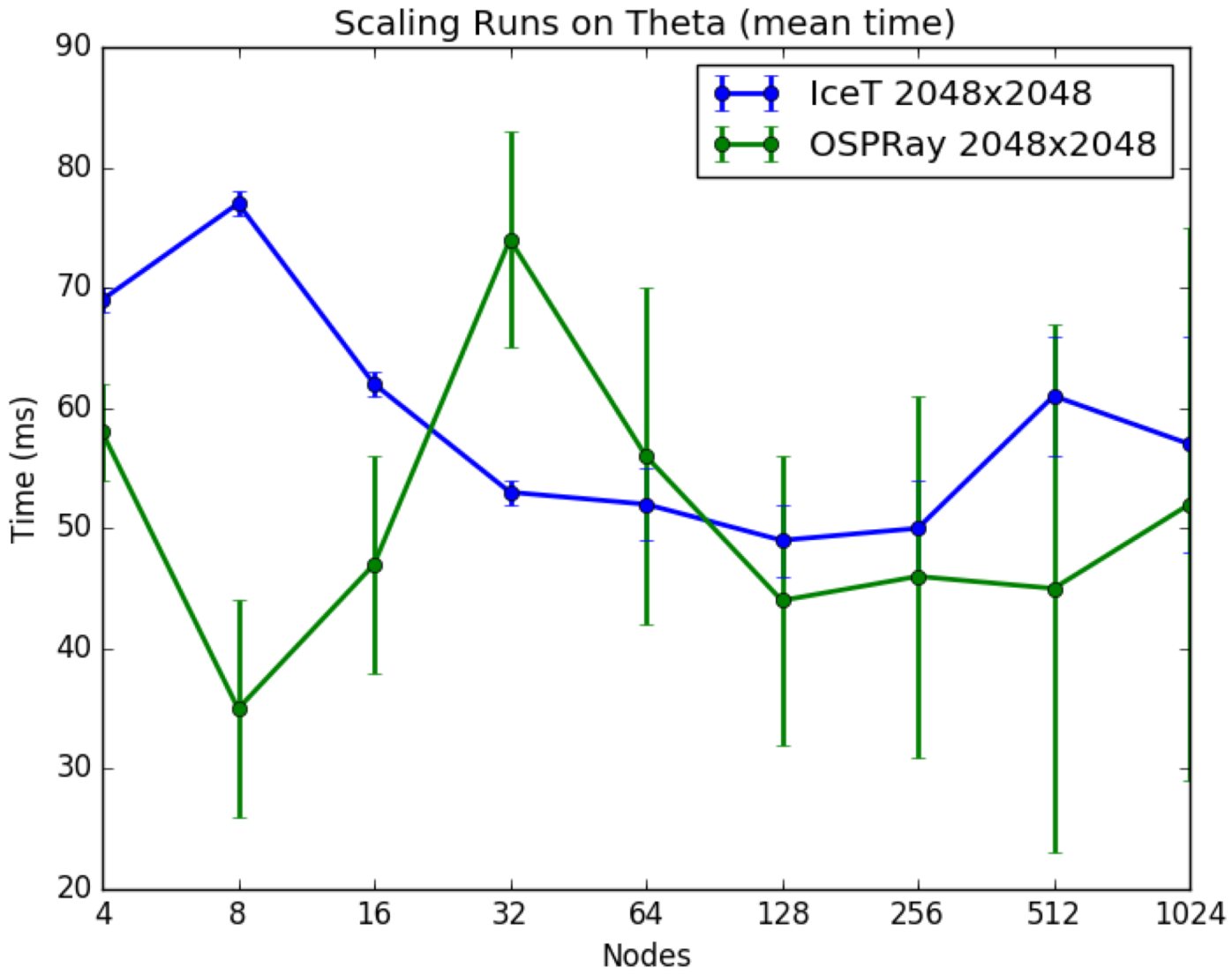


TACC

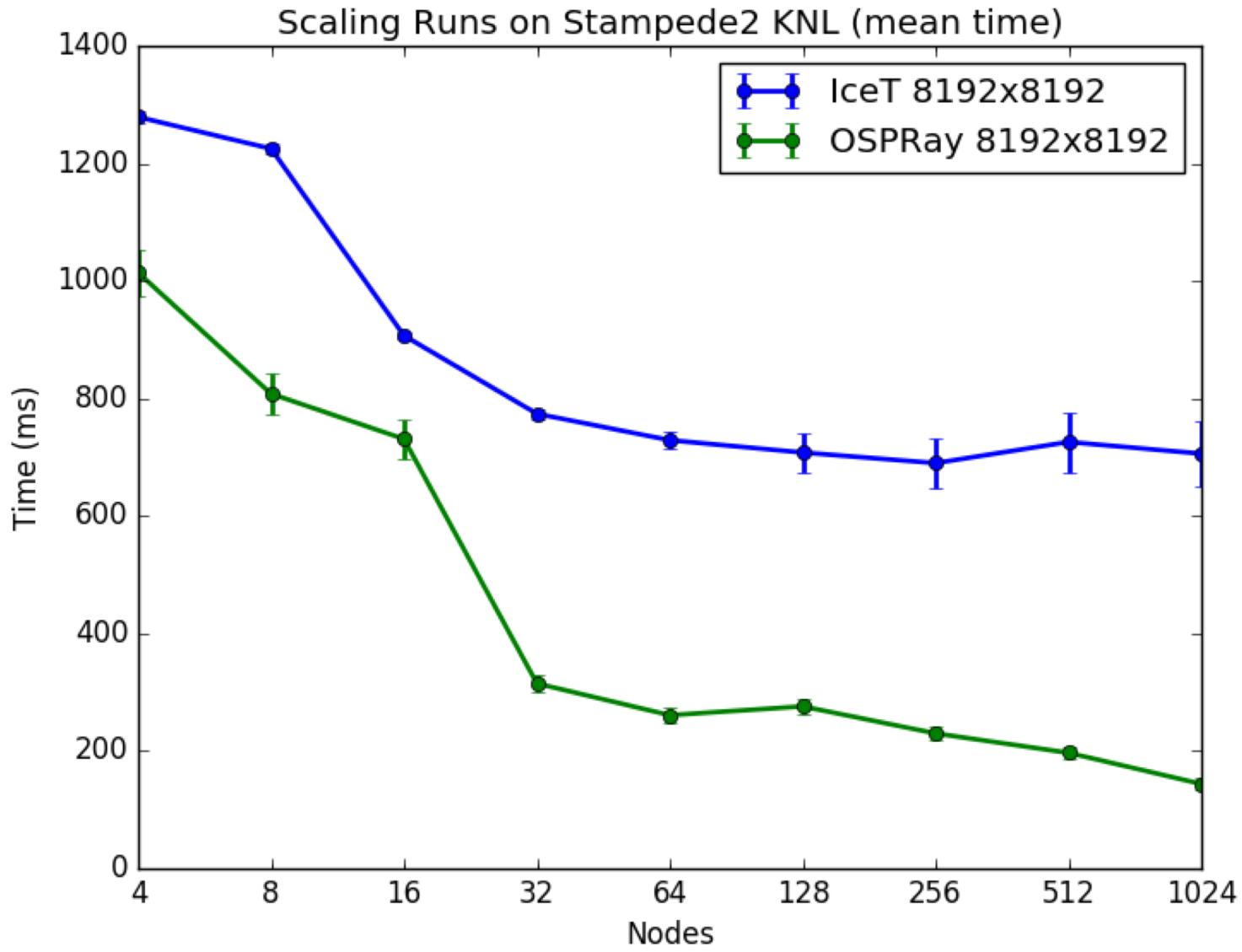
Stampede2 SKX – 2K x 2K



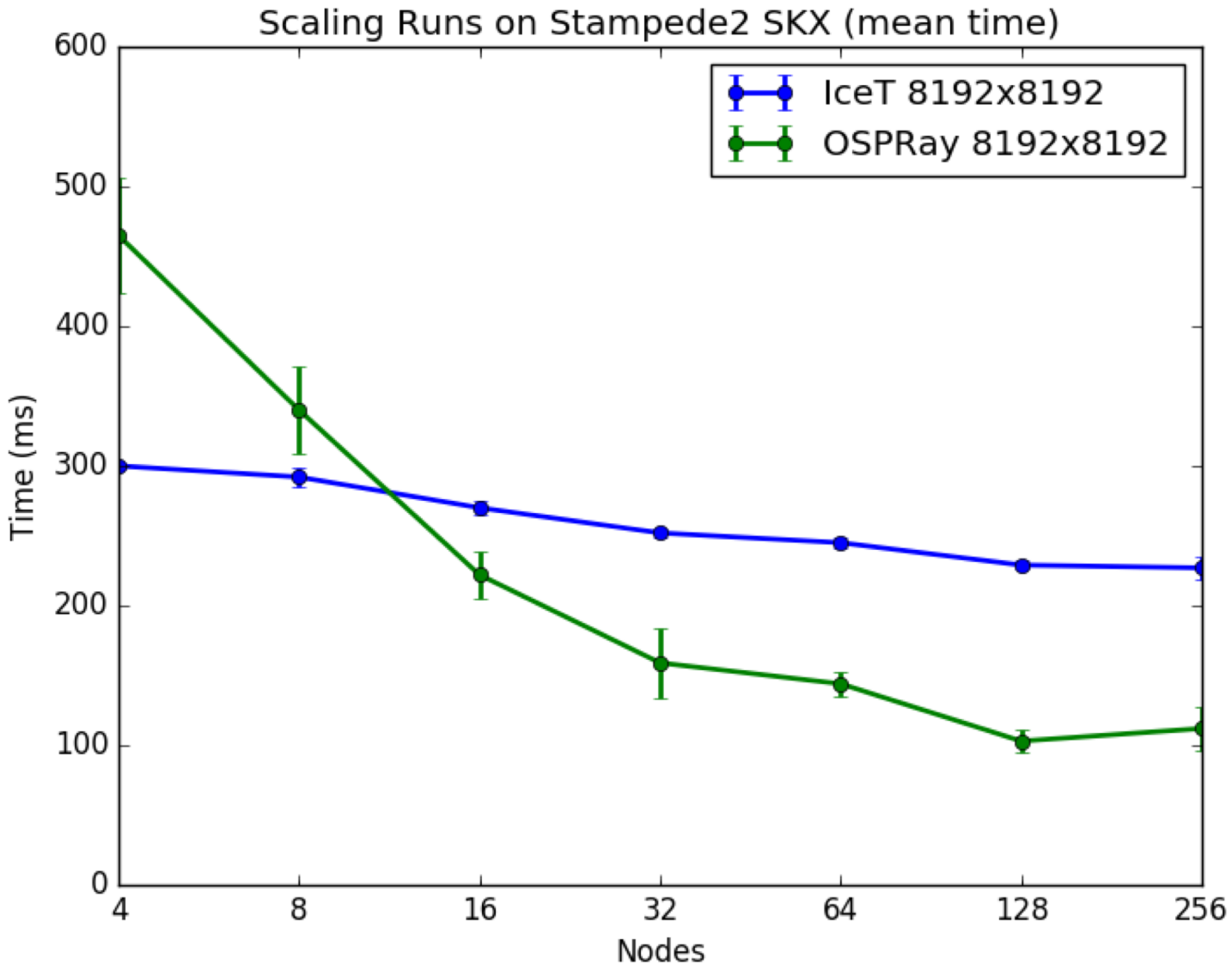
Theta KNL – 2K x 2K



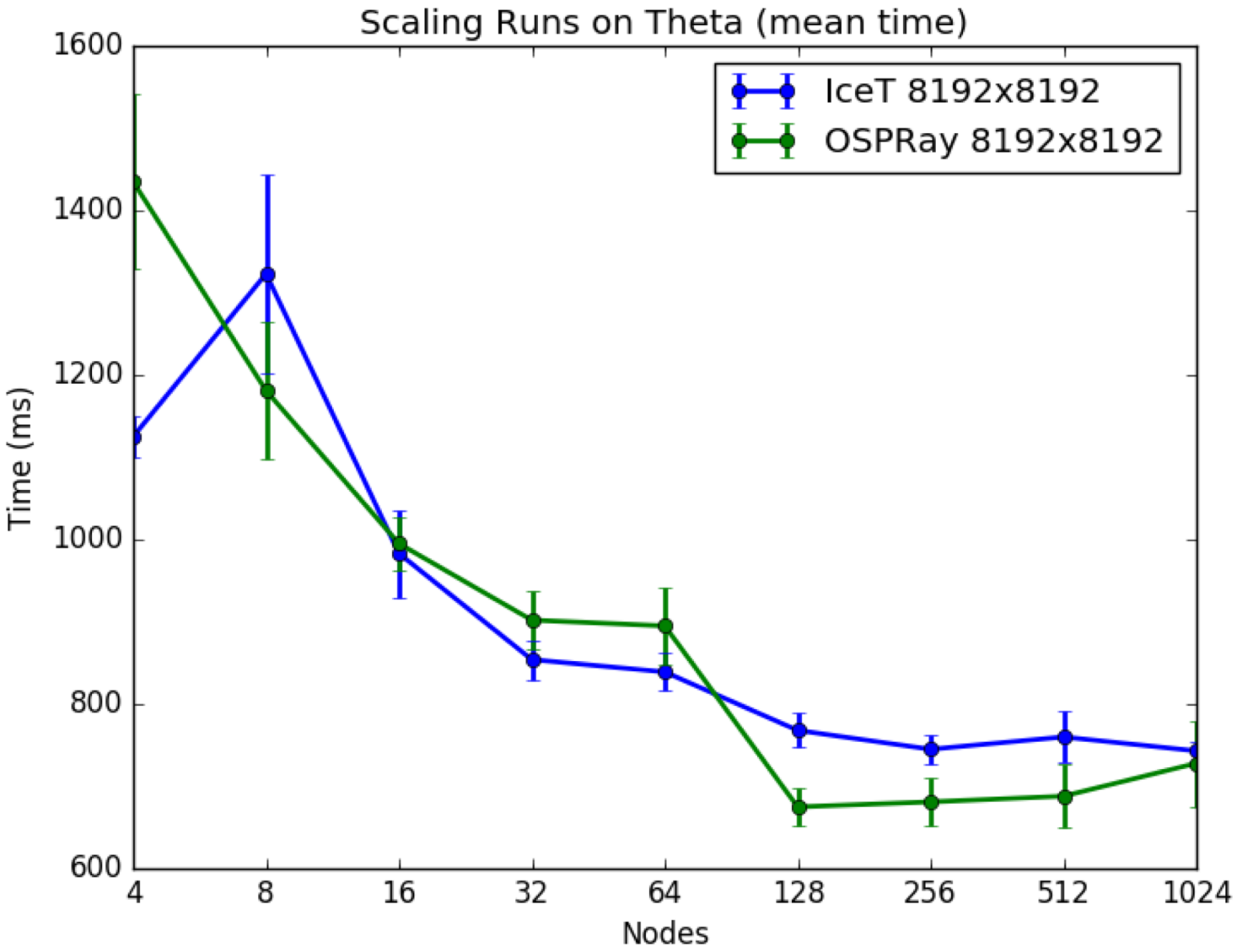
Stampede2 KNL – 8K x 8K



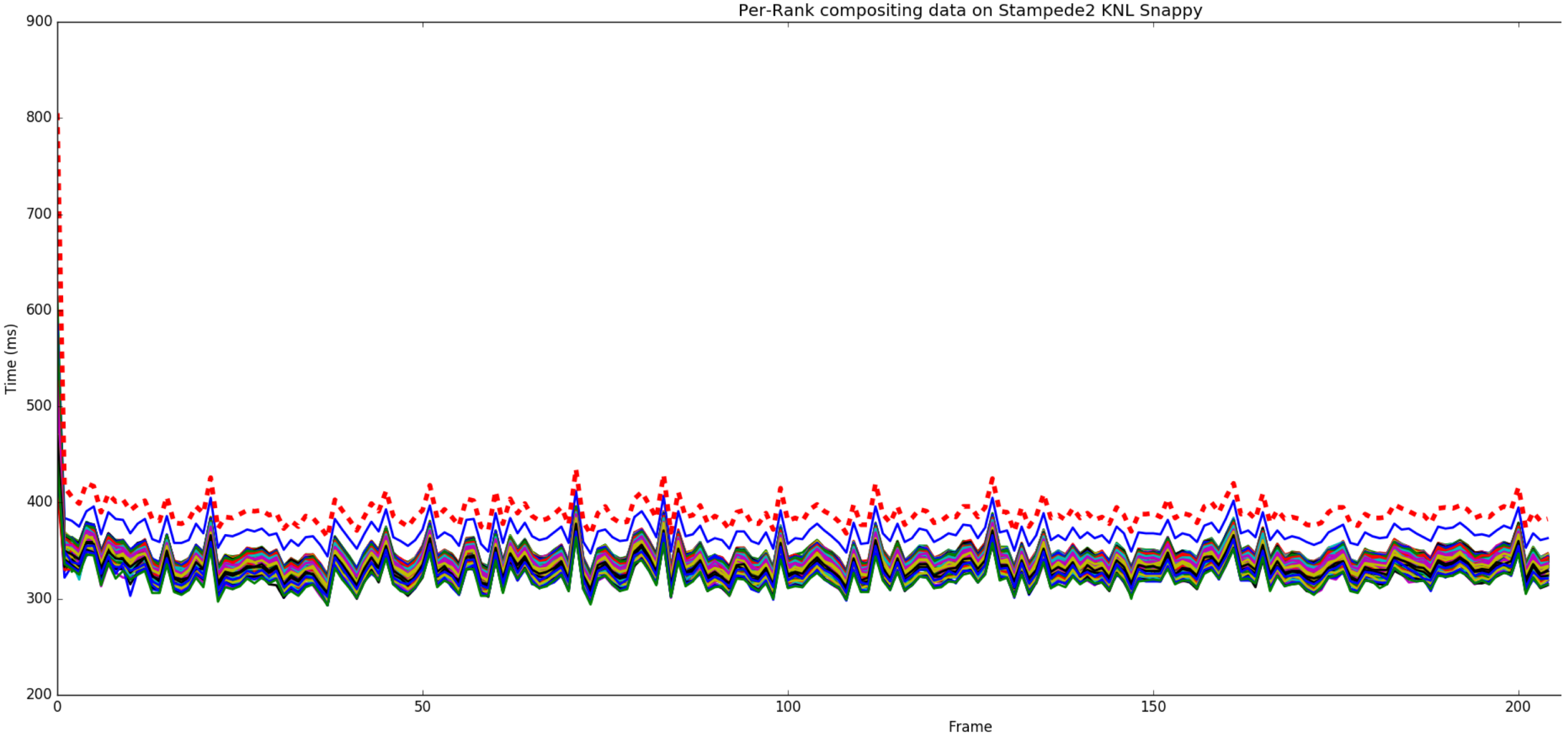
Stampede2 SKX – 8K x 8K



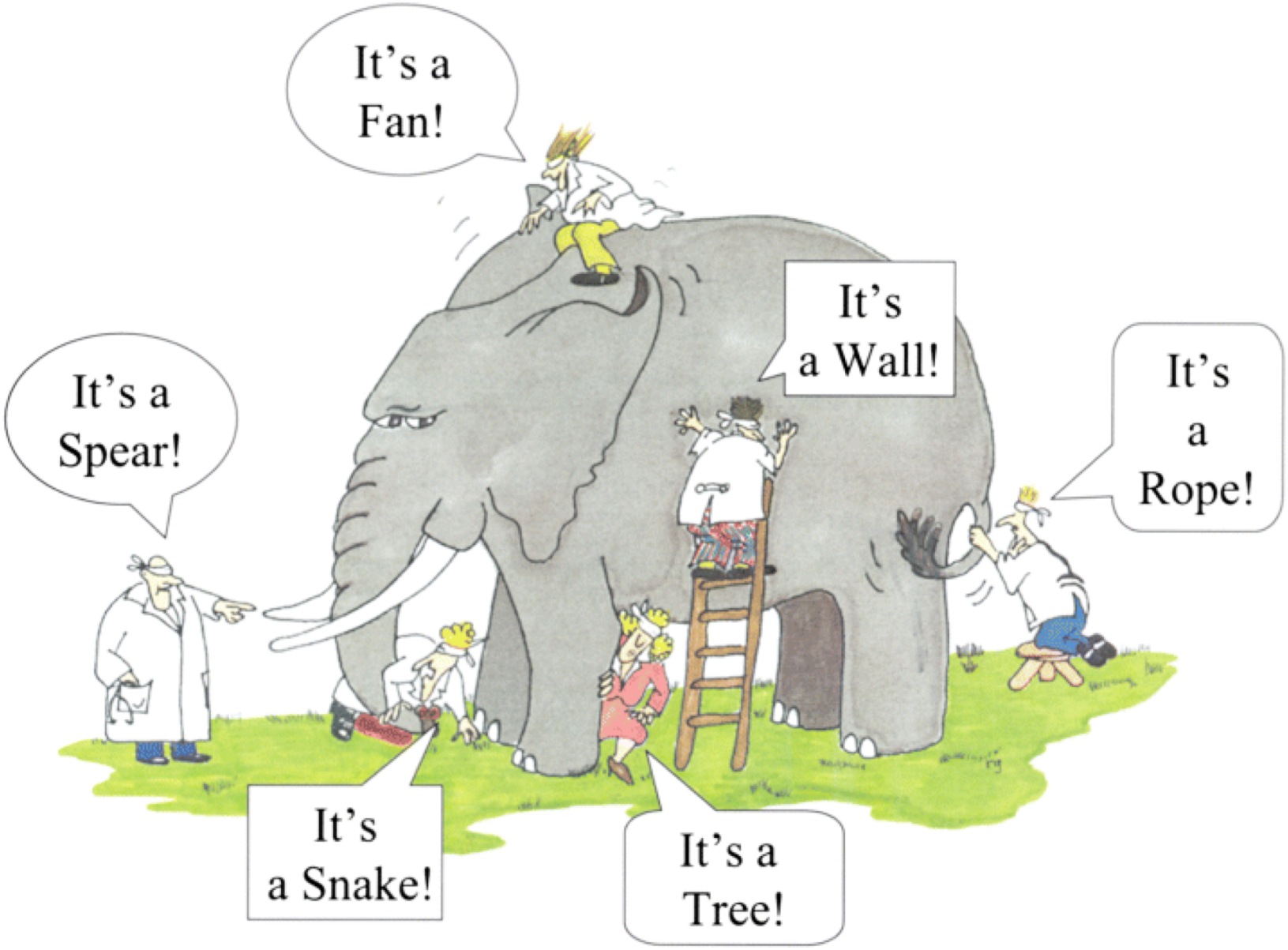
Theta KNL – 8K x 8K



Stampede2 KNL – Compositing per rank
128 nodes @ 8K x 8K



Discussion



Thank you!

pnav@tacc.utexas.edu