System Sotware Research (2)

Towards Next-Generation Supercomputing

Fault tolerant Infrastructure for Billion-Way Parallelization

Asynchronous Checkpointing System

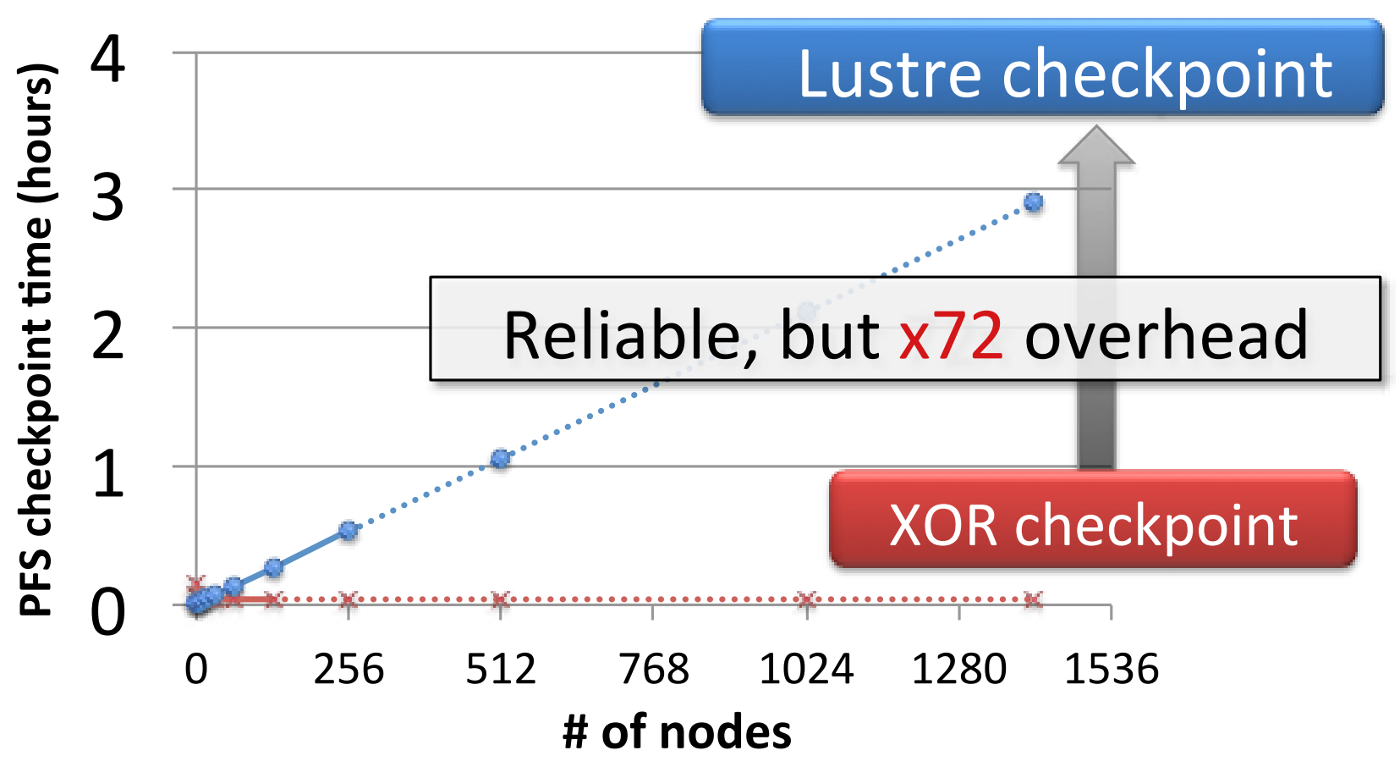
Background: Increasing system failures

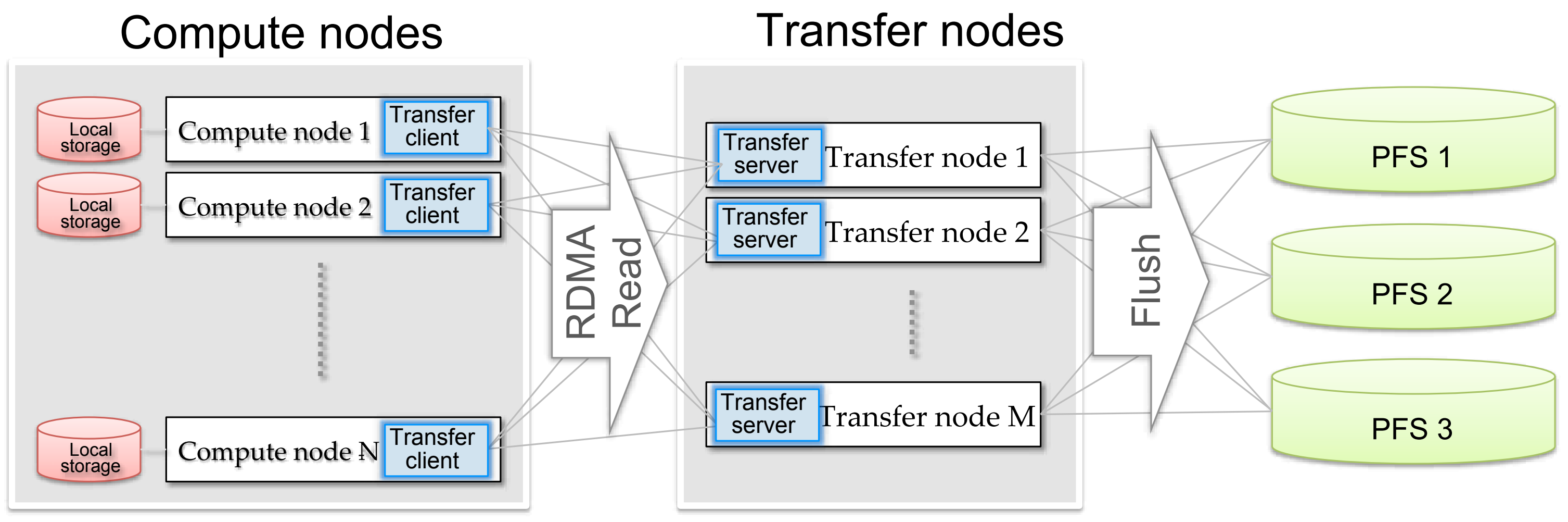
- A node failure occurred every 13 hours on average
- A parallel file system (PFS) checkpointing overhead (3 hours)

Objective : Reduce PFS checkpoint overhead

Proposed method : Implementation and modeling of an asynchronous checkpointing

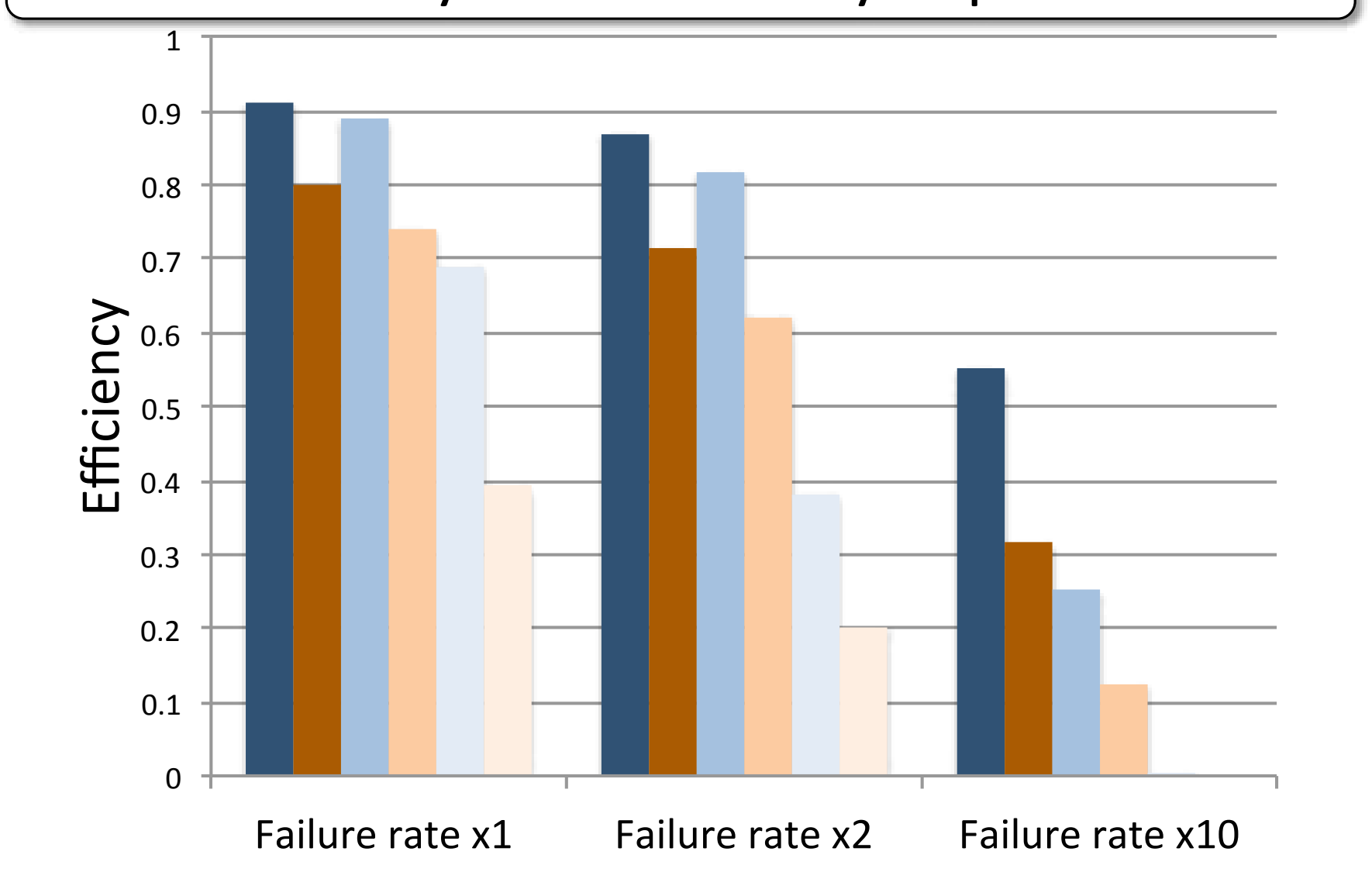
TSUBAME2.0 Lustre checkpoint time





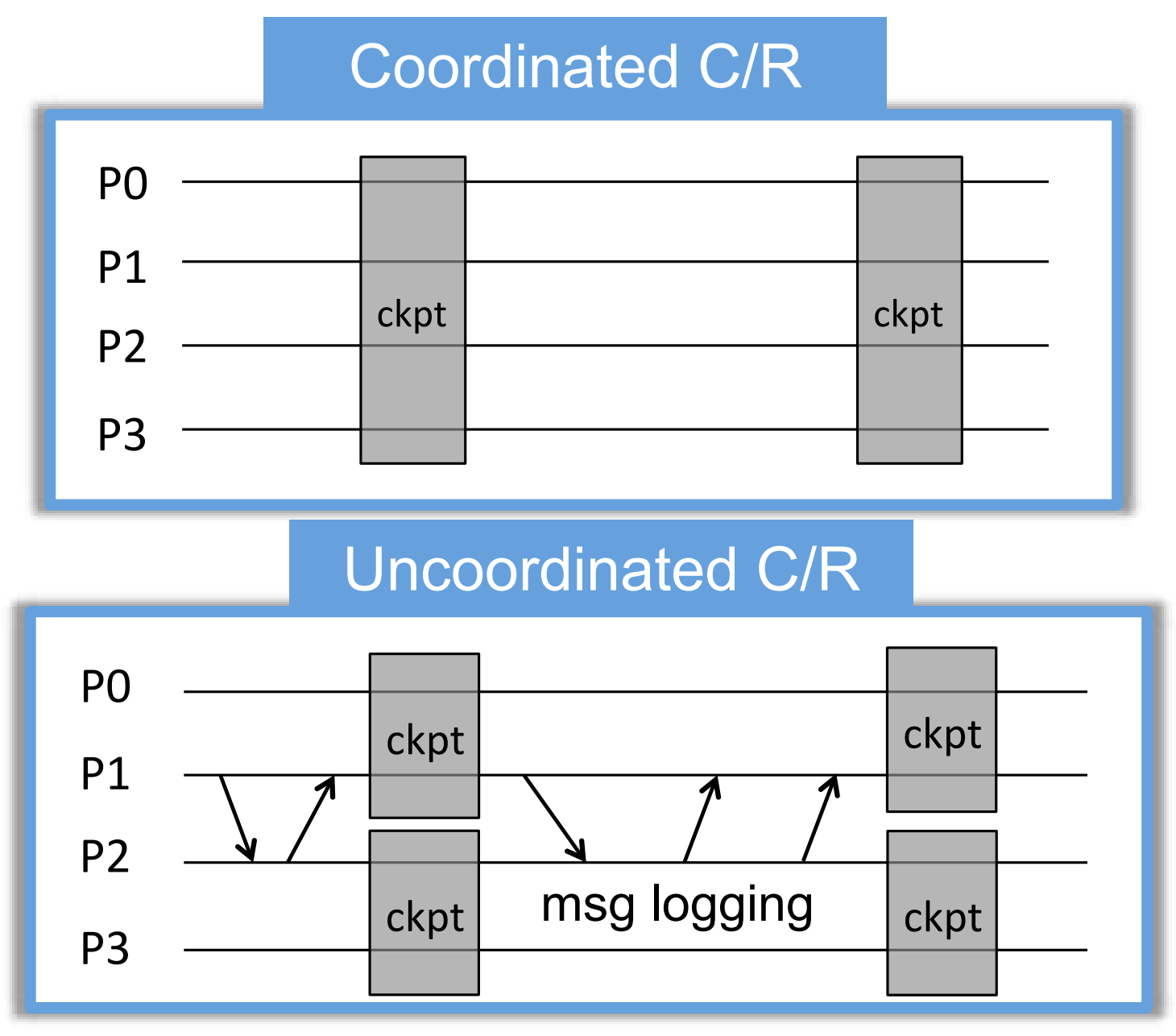
x1.1 ~ x1.8 System efficiency improvement

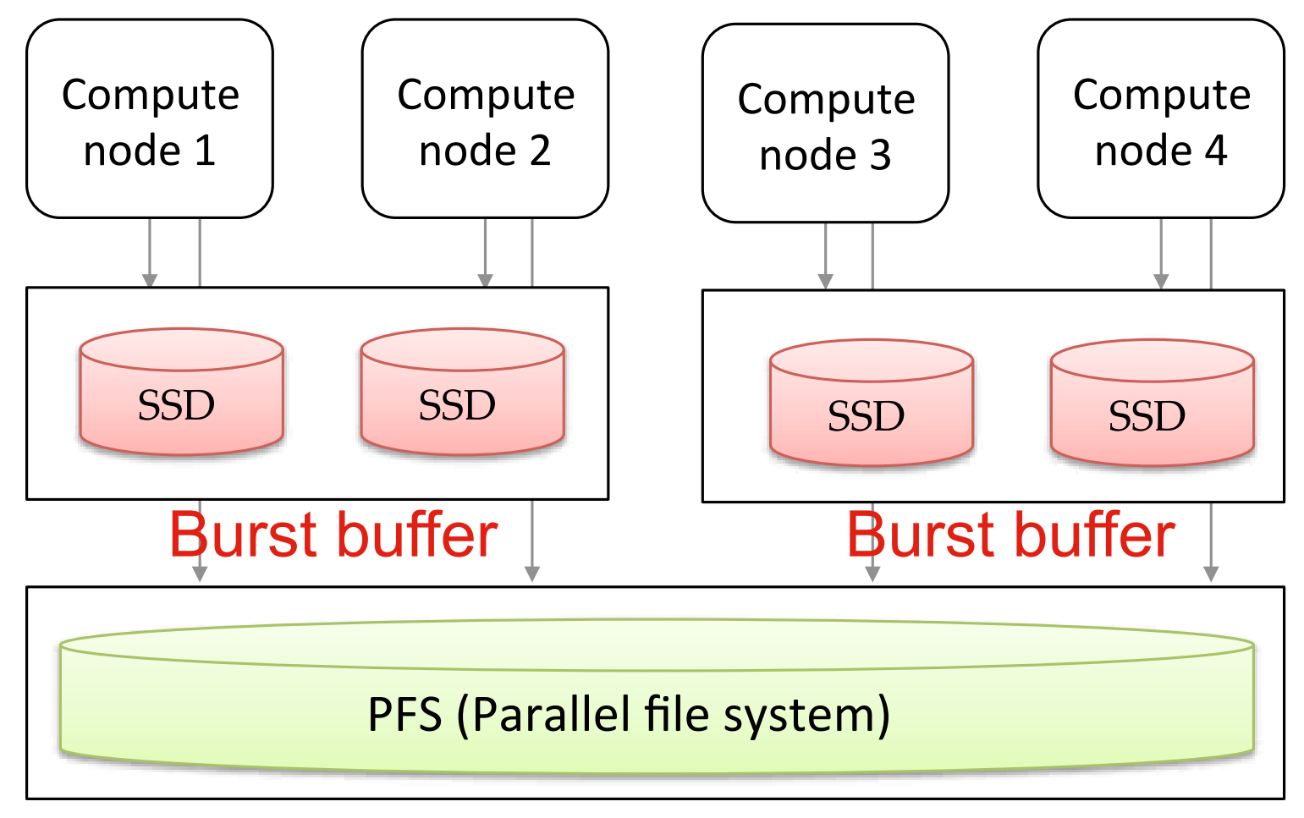
$$\text{Efficiency} = \frac{\text{ideal\_runtime}}{\text{expected\_runtime}}$$

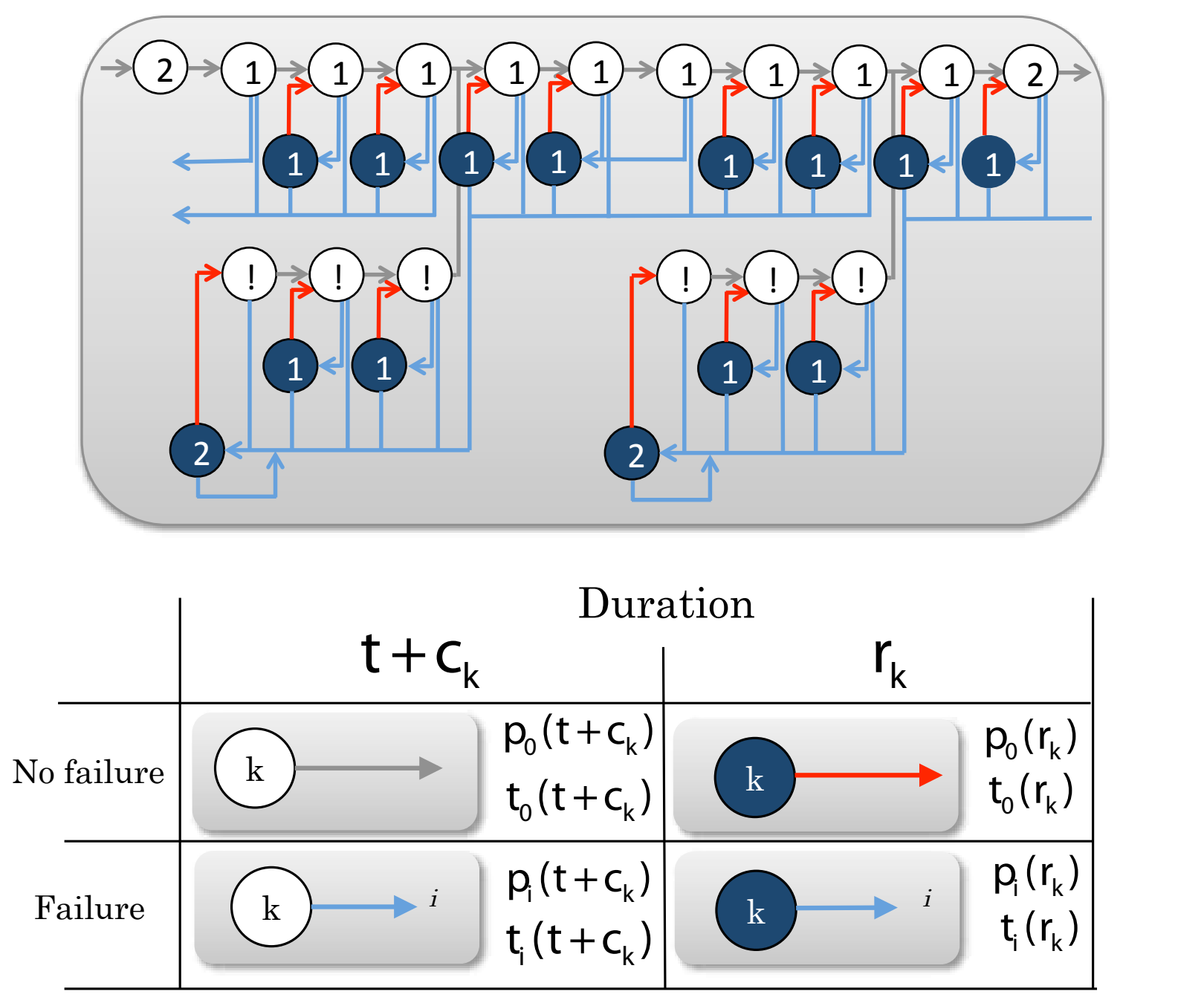


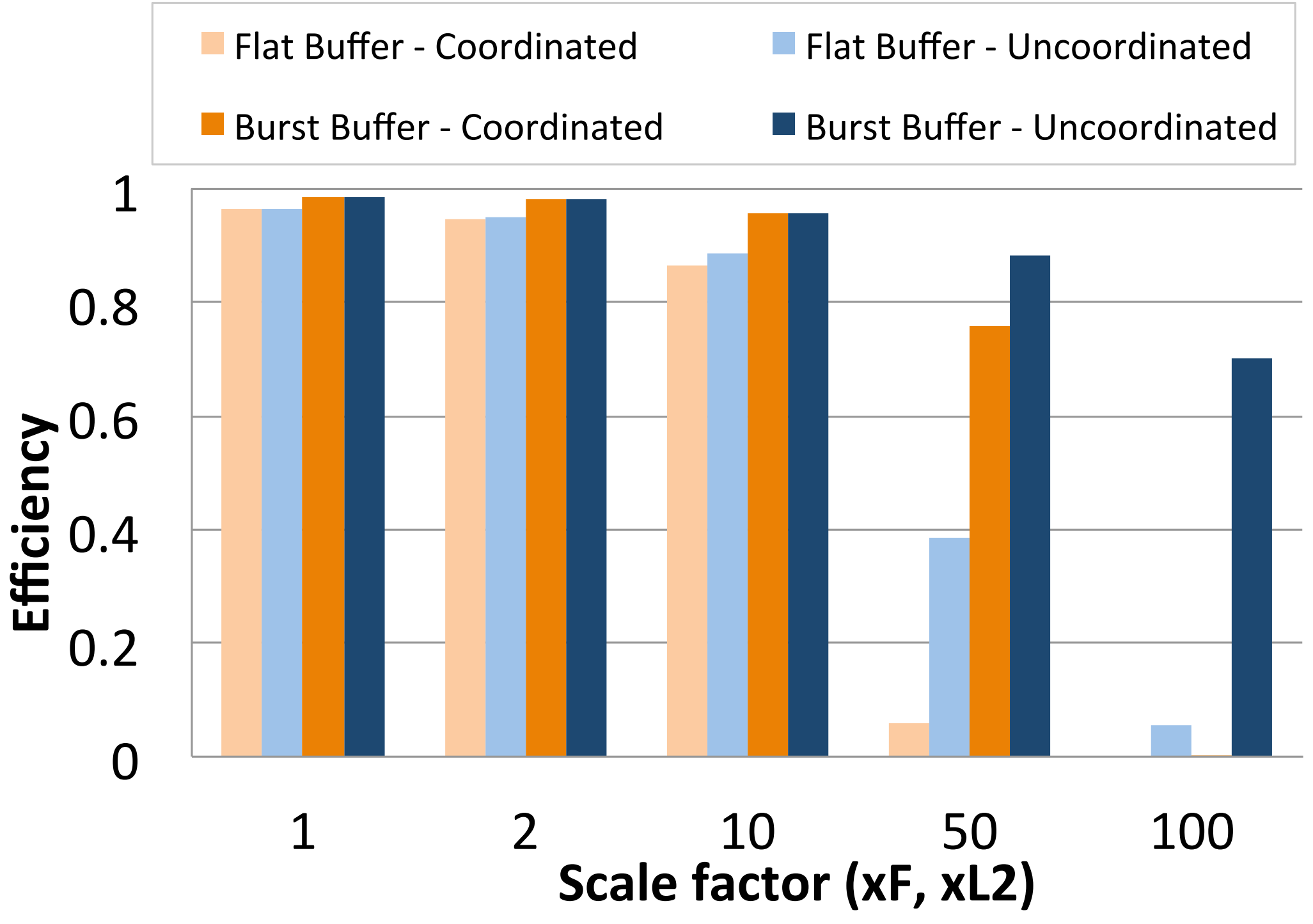
Multi-tier Resilient Storage Design

- A burst buffer is a storage space to bridge the gap in latency and bandwidth between node-local storage and the PFS
  - Shared by a subset of compute nodes
- Although additional nodes are required, several advantages
  - More Reliable because burst buffers are located on a smaller # of nodes
  - Efficient utilization of storage resources with uncoordinated checkpointing









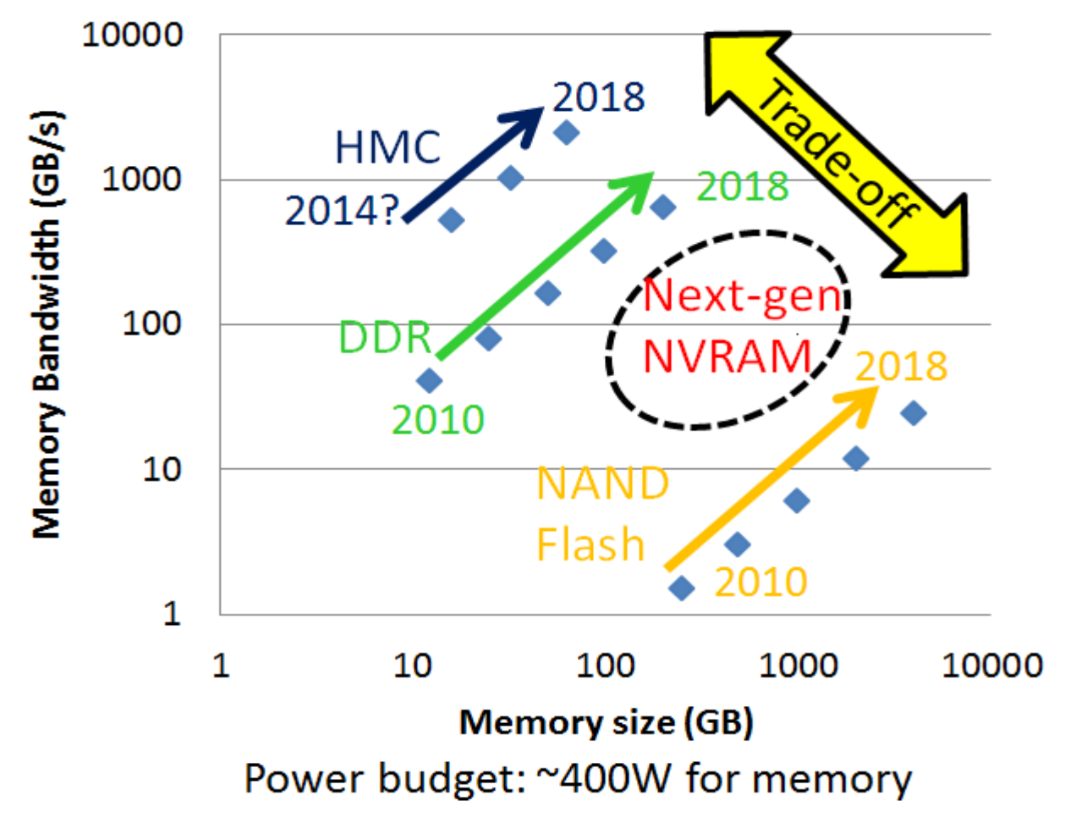
This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52- 07NA27344. LLNL-POST-644954

Dealing with Deeper Memory Hierarchy

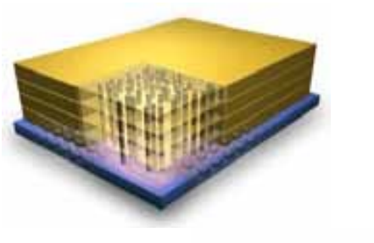
In Exa-scale supercomputing systems, the "**memory wall**" problem will become even higher, which prevents the realization of exa-scale real world simulations.
In our project, "Software Technology that Deals with Deeper Memory Hierarchy in Post-petascale Era", we promote research in aspect of "Architecture", "Algorithm" and "System software".

[Architecture]

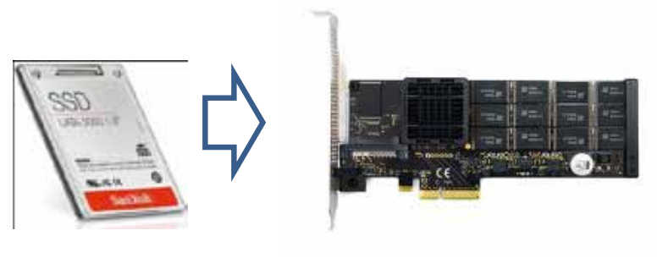
To suppose supercomputerm architecture with **deeper memory hierarchy** including hybrid memory devices, including non-volatile RAM (NVRAM).



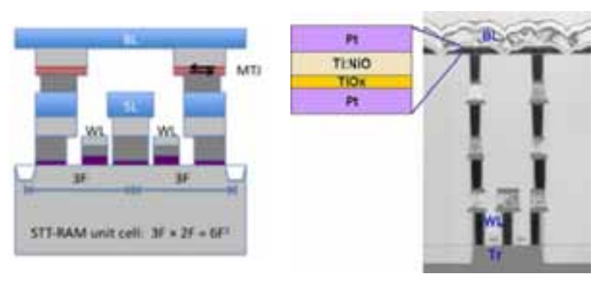
Power budget: ~400W for memory



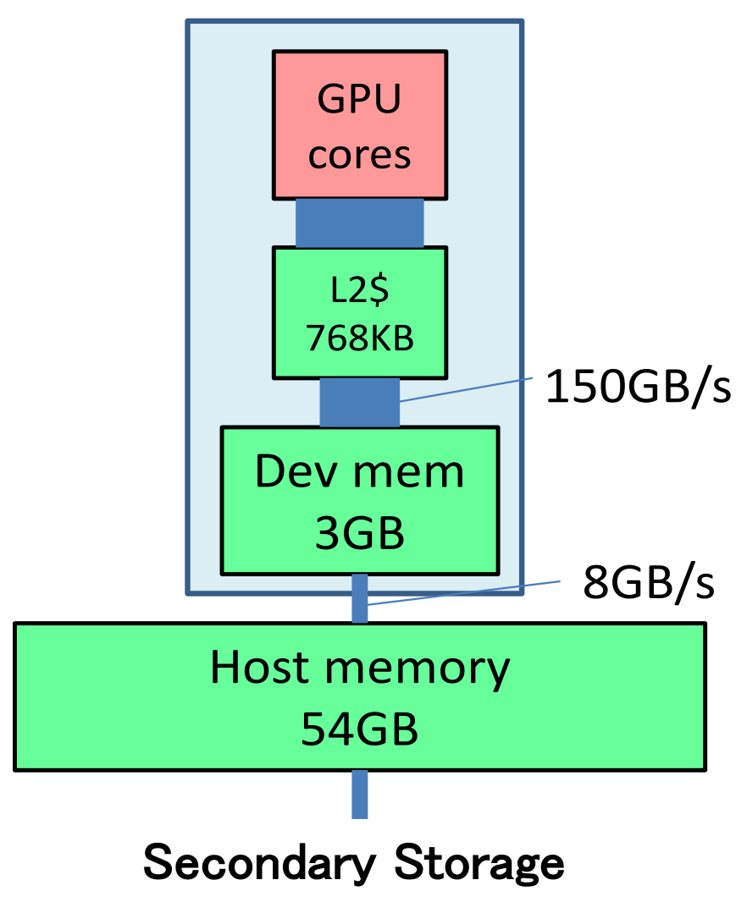
Hybrid Memory Cube (HMC):
DRAM chips are stacked with TSV technology.
It will have advantage in bandwidth over DDR, but capacity will be smaller.



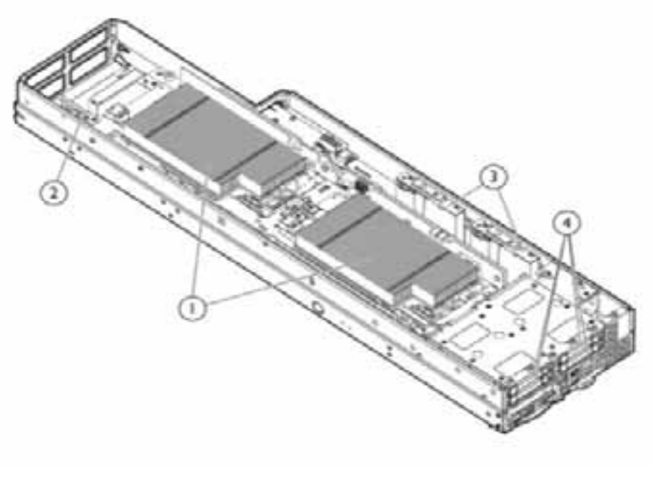
NAND Flash:
SSDs are already commodity.
Newer products, such as IO-drive have O(GB/s) bandwidth.



Next-gen non-volatile RAM (NVRAM):
Several kinds of NVRAM such as STT-MRAM, ReRAM, FeRAM, etc, will be available in a few years.

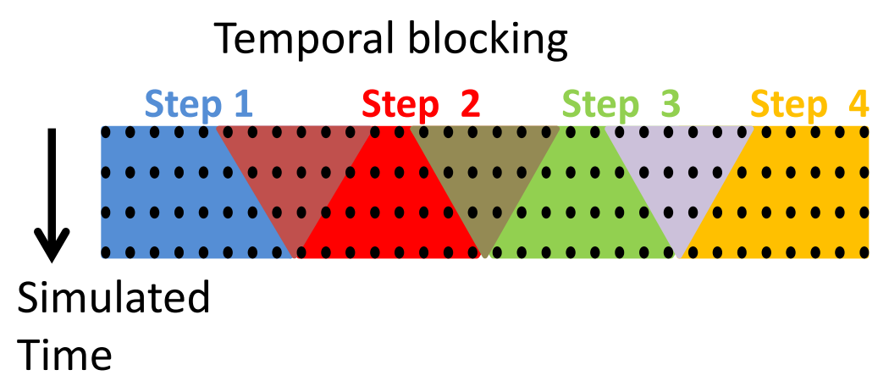


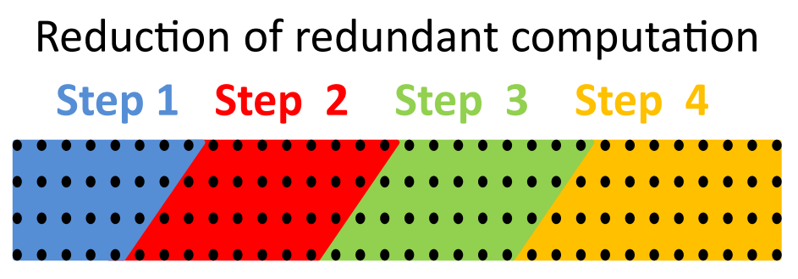
Currently, we use TSUBAME2, CPU-GPU hybrid supercomputer as research environment.
Here we have memory hierarchy of GPU device memory and Host memory.



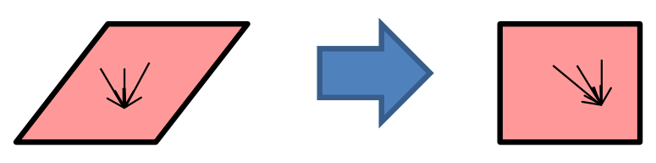
[Algorithm]

To harness hierarchical memory efficiently, we are investigating **locality improvement** of application algorithms.
In stencil applications, **temporal blocking** is the key.

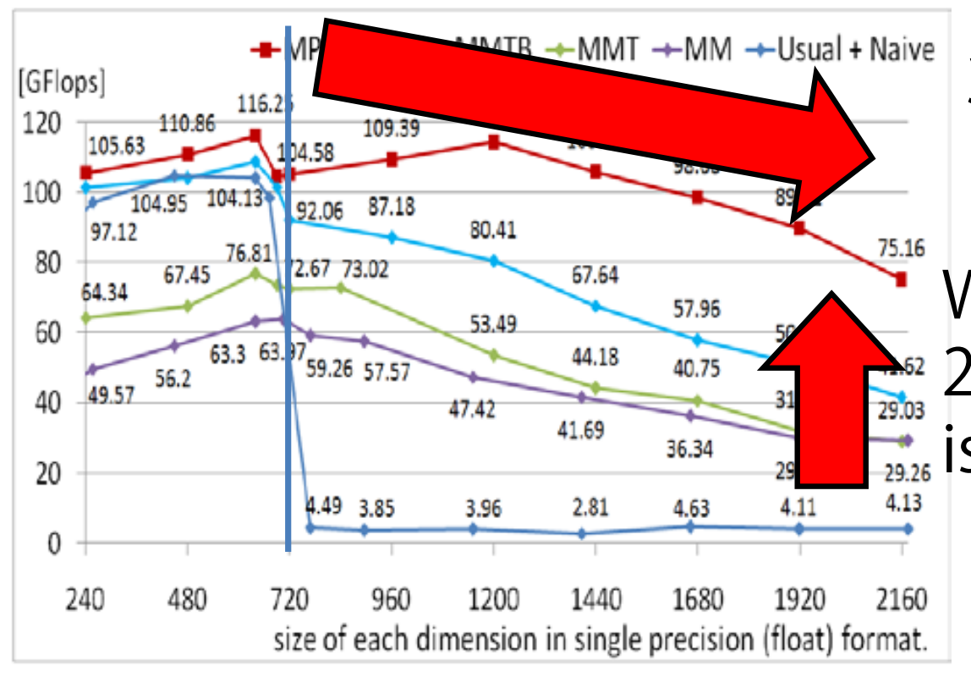




Cf) James Demmel et al.

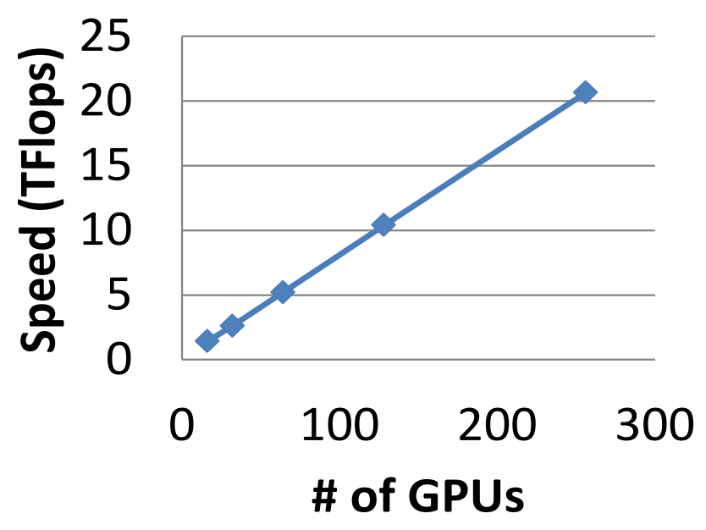


Reduction of buffer utilization [AsHES 13]



3D 7-point stencil on a M2050 GPU

With optimized temporal blocking, 27 tiems larger array than GPU memory is efficiently used (only 30% overhead)!



Good weak scalability

[System Software]

To support real applications to harness hierarchical memory with lower development efforts, system software support is necessary. Our target includes **locality aware compiler** and **scalable memory management runtime**.

PI: Toshio Endo. Supported by JST-CREST