Evaluation of Parallel I/O Performance and Energy with Frequency Scaling on Cray XC30

Suren Byna and Brian Austin

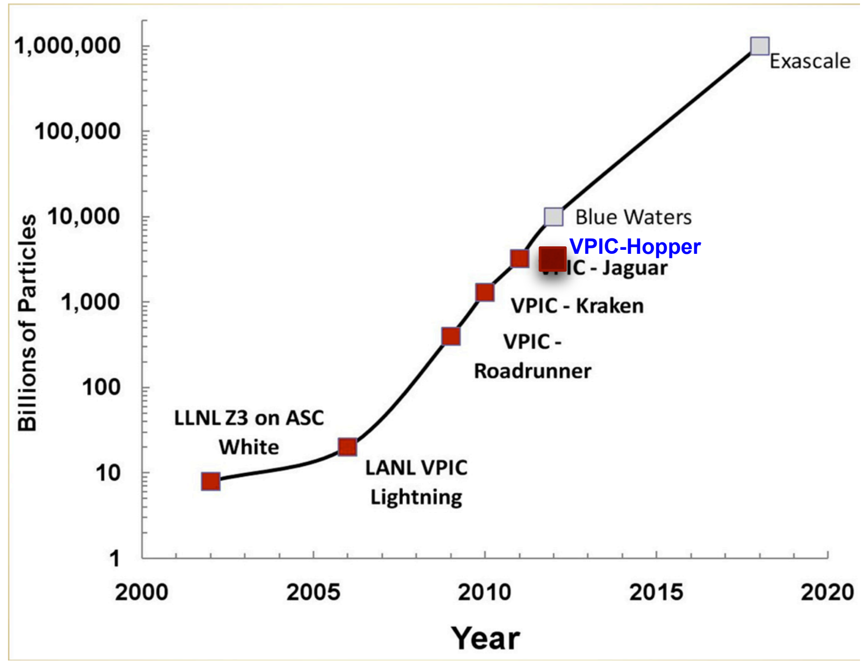
Lawrence Berkeley National Laboratory

Energy efficiency at Exascale

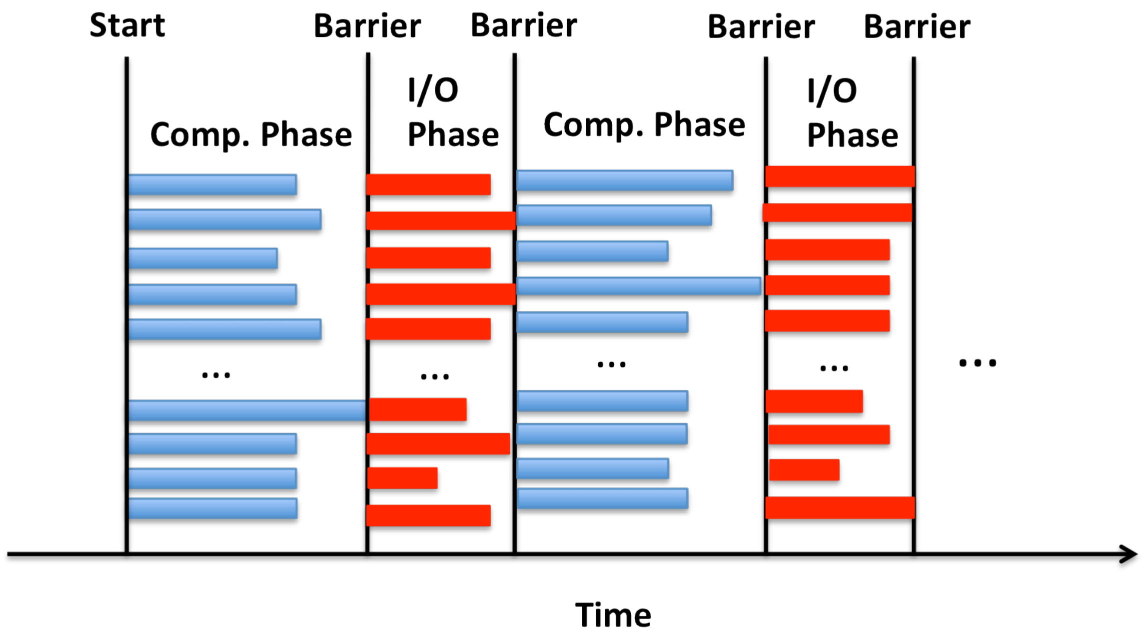
- A design goal for future exascale systems – Power consumption less than 20 MW
- Dynamic voltage and frequency scaling (DVFS) is a method to provide variable amount of energy on processors
- Lowering frequency saves CPU power
- When CPUs have to do computing in low power states, DVFS may affect performance
- Several efforts to avoid or reduce performance degradation with DVFS

Large-scale Scientific Simulations

- Large-scale scientific simulations use significant portion of supercomputers
  - VPIC
  - Flash
- Produce large amounts of data
  - 10 trillion particles – 8 properties: ~300 TB



DVFS during I/O phases



- Simulations with interleaving computation and I/O phases
- Parallel I/O is often a collective operation in writing to a single file
- If processors are idle or not computing during I/O, can they be in a low power state?
- **What is the impact of DVFS during I/O phases?**

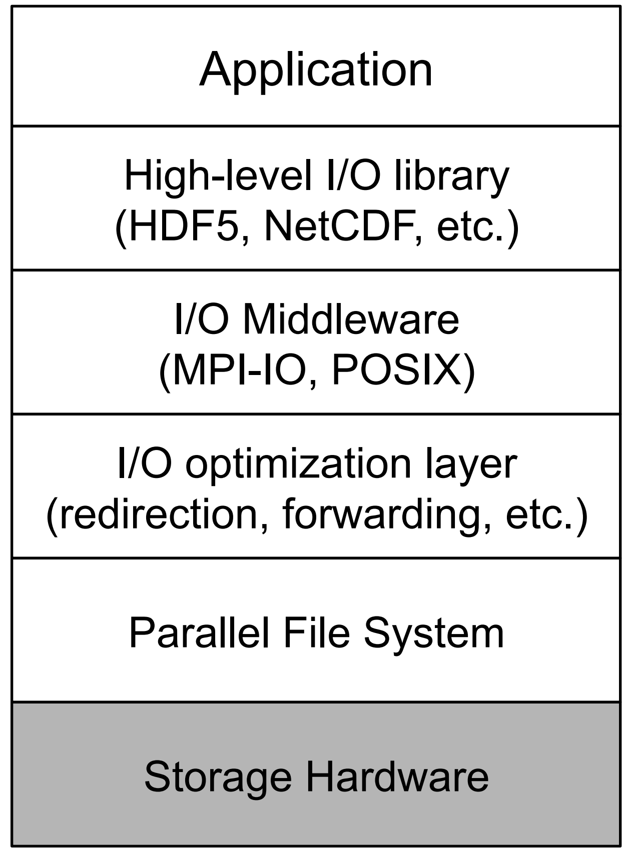
Parallel I/O

Parallel I/O software stack

- Application
- High-level I/O libraries and data models
- I/O middleware
- Parallel file system

Options for performance optimization

Complex inter-dependencies among layers



Experimental Setup

- NERSC Edison
  - Cray XC30
  - Each node has two 2.4 GHz 12-core Intel Ivy Bridge CPUs
  - 64 GB DDR3 DRAM
  - Cray Aries interconnect
- File system
  - Sonexion 1600 appliance w/ Lustre
  - 144 OSTs with 72 GB/s peak I/O bandwidth
  - 32 MB stripe size

Applications

- VPIC-IO – I/O kernel from a plasma physics simulation
  - VPIC is developed at LANL and I/O w/ HDF5 at LBNL
  - H5Part and HDF5 I/O
  - Each MPI process writes data for 8 million particles
  - Each particle has 8 properties
  - Each property is stored as a 1D HDF5 dataset
- VORPAL-IO – I/O kernel from accelerator physics
  - Developed at TechX
  - I/O kernel extracted at LBNL
  - H5Block and HDF5 I/O
  - Each process writes a 3D block of 100 x 100 x 60

Measurements

- I/O time
  - Maximum time of all processes
  - Includes file open, close, and write times
  - Ran each experiment 5 times and selected the best performing
- Energy and Power measurements
  - Cray Power Management counters – PM counters
  - /sys/cray/pm_counters/cpu_{energy,power}
  - Developed a small library to obtain the elapsed power/energy and to aggregate from all nodes involved in running a job – PMON
  - To set frequency for running an I/O kernel

    e.g.: `aprun --p-state=1800000 -n 2048 exec args`

The power and energy measurements are for compute nodes, not for the I/O subsystem

Scaling tests

Weak Scaling

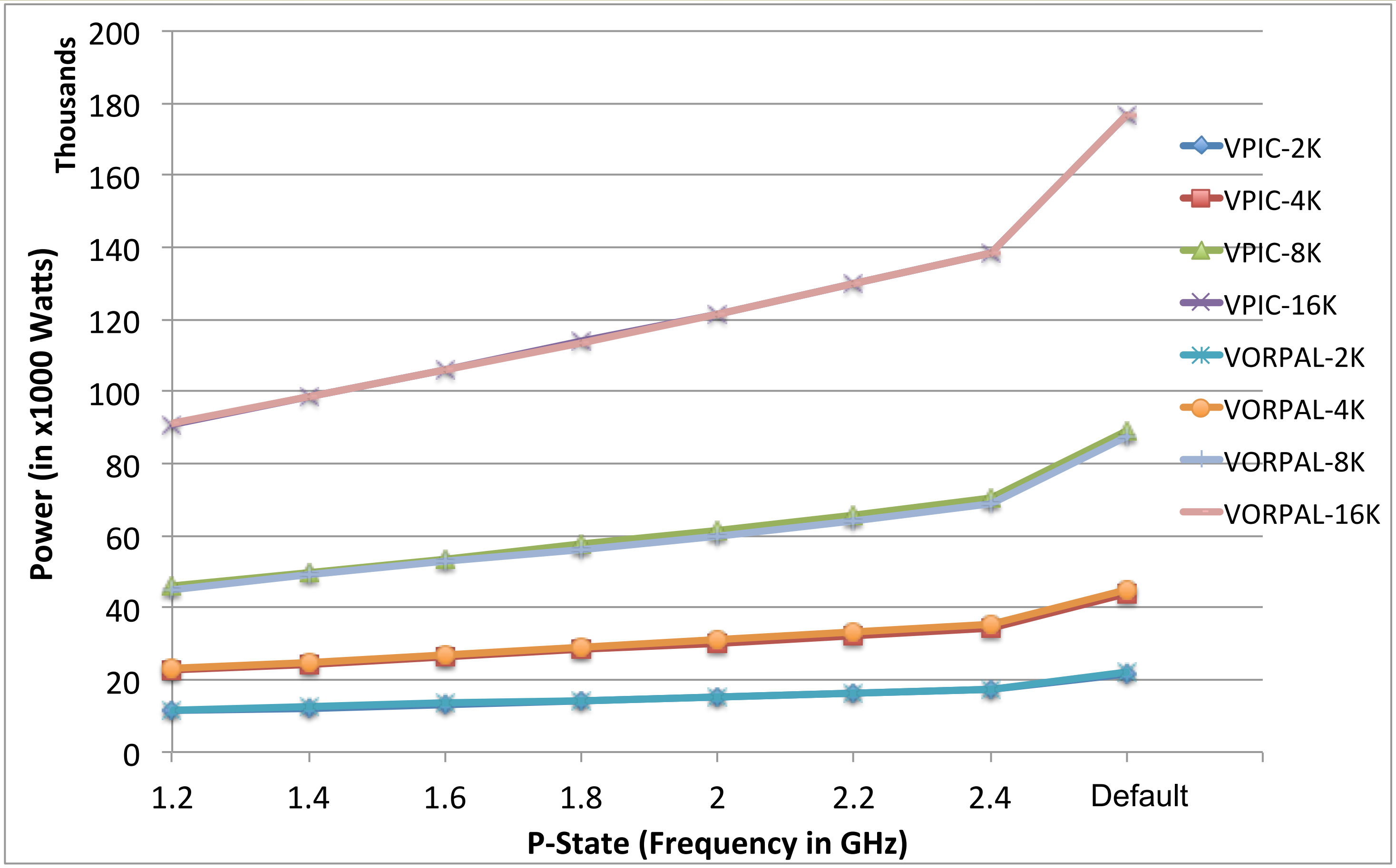
Number of cores	VPIC data size	VORPAL data size
2K	512 GB	1.6 TB
4K	1 TB	3.2 TB
8K	2 TB	6.4 TB
16K	4 TB	12.8 TB

Strong Scaling

Number of cores	VPIC data size
2K	1 TB
4K	
8K	

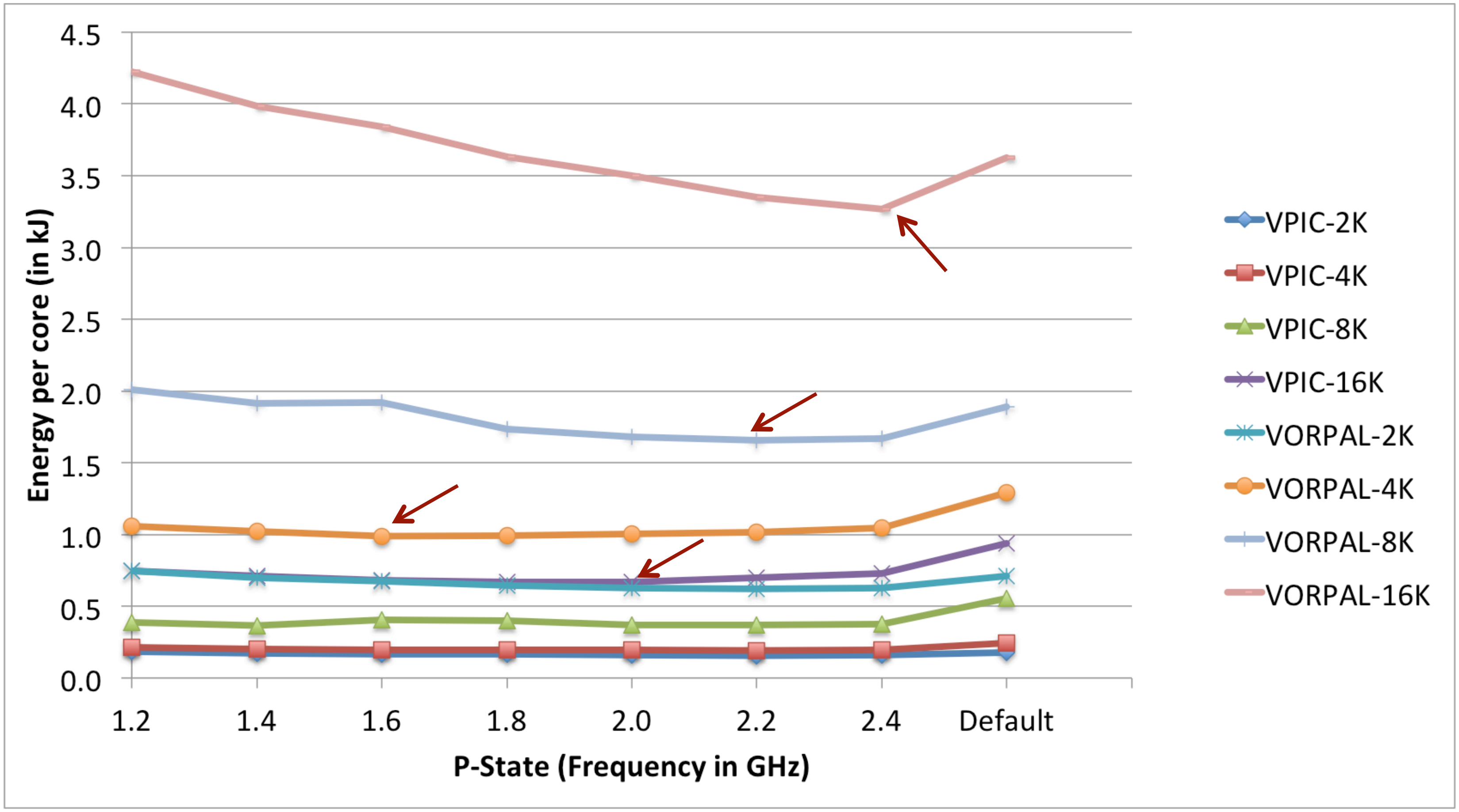
Weak-scaling – Power Consumption

Better



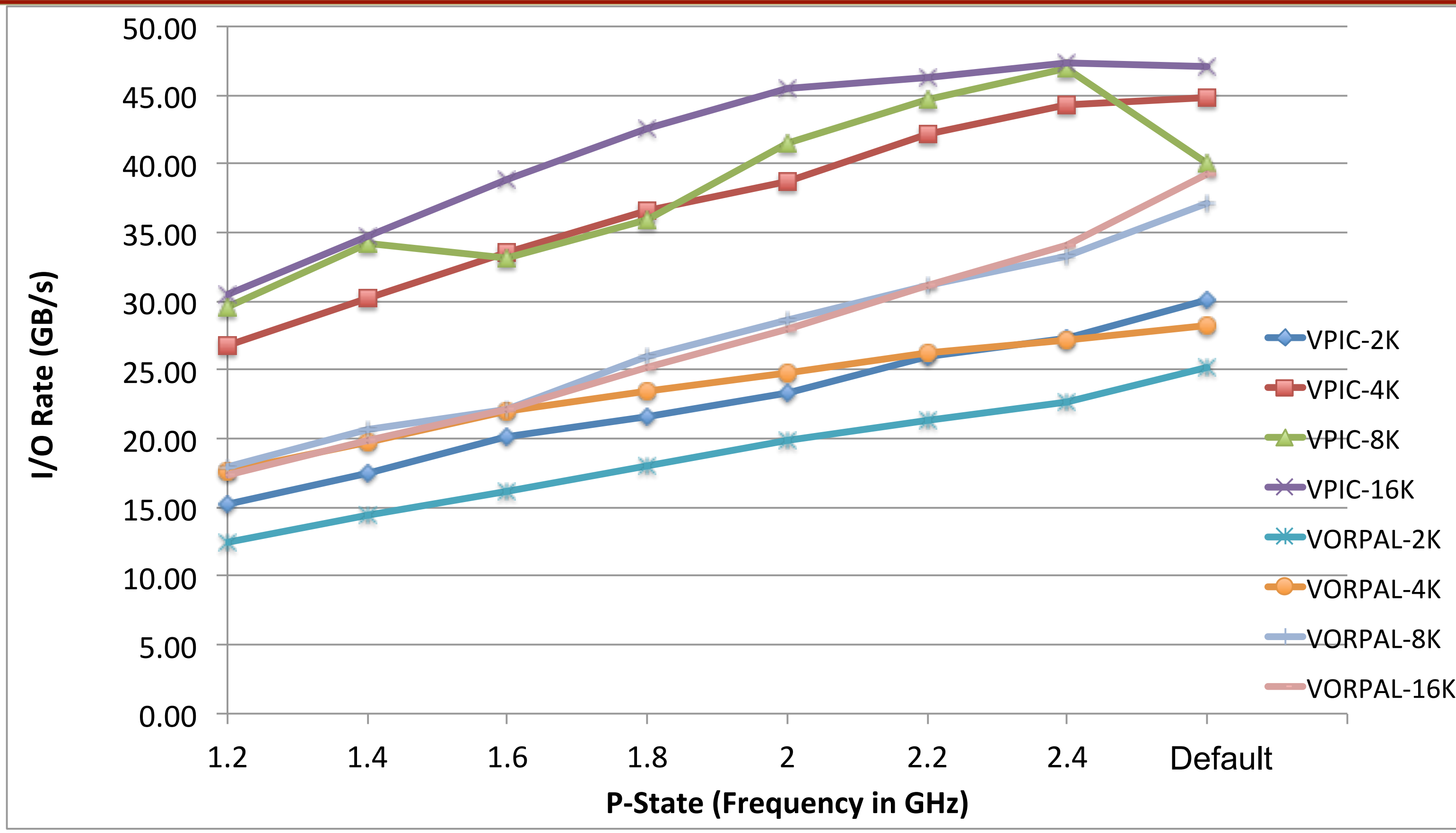
Weak-scaling – Energy Consumption

Better



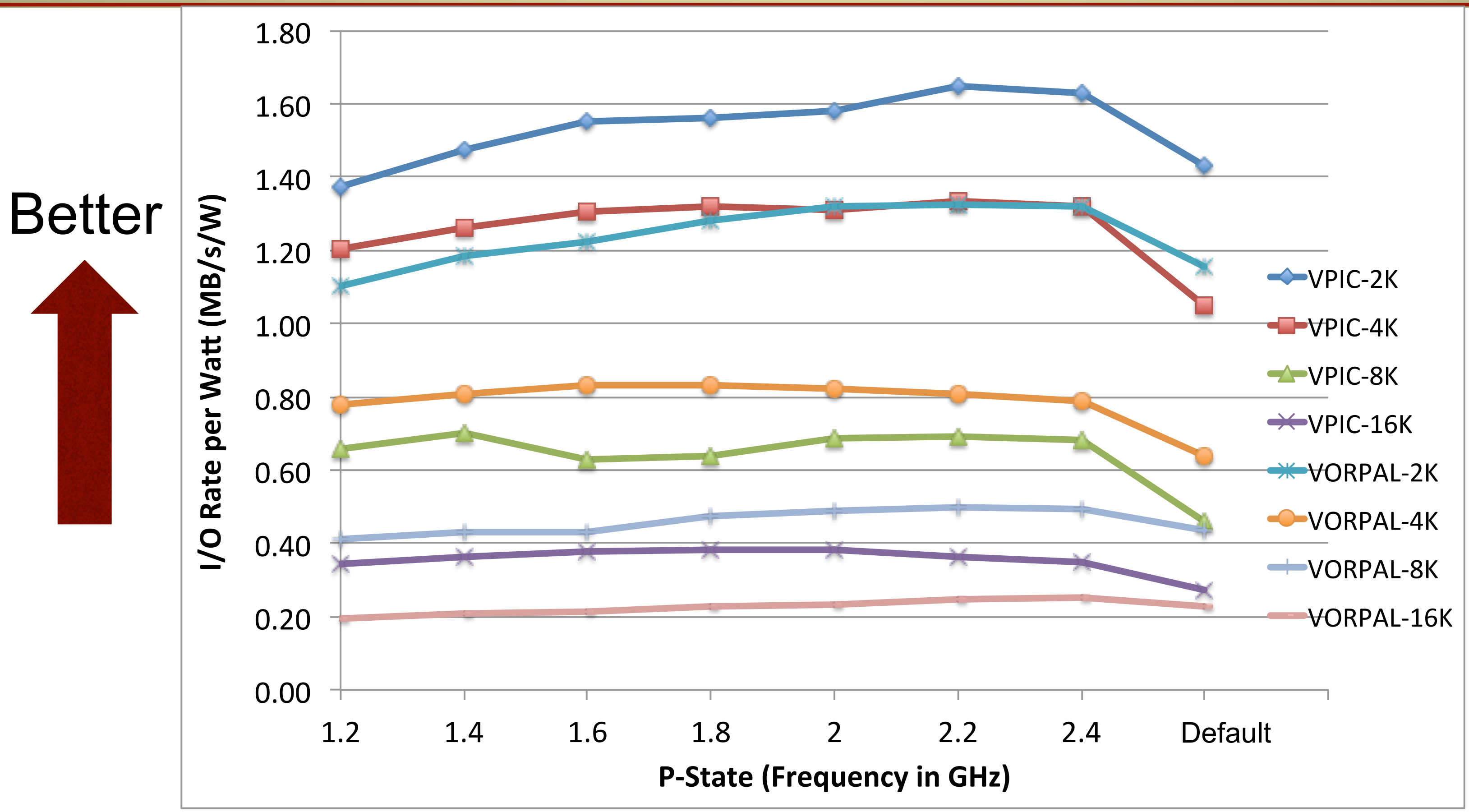
Weak-scaling – I/O Rate

Better

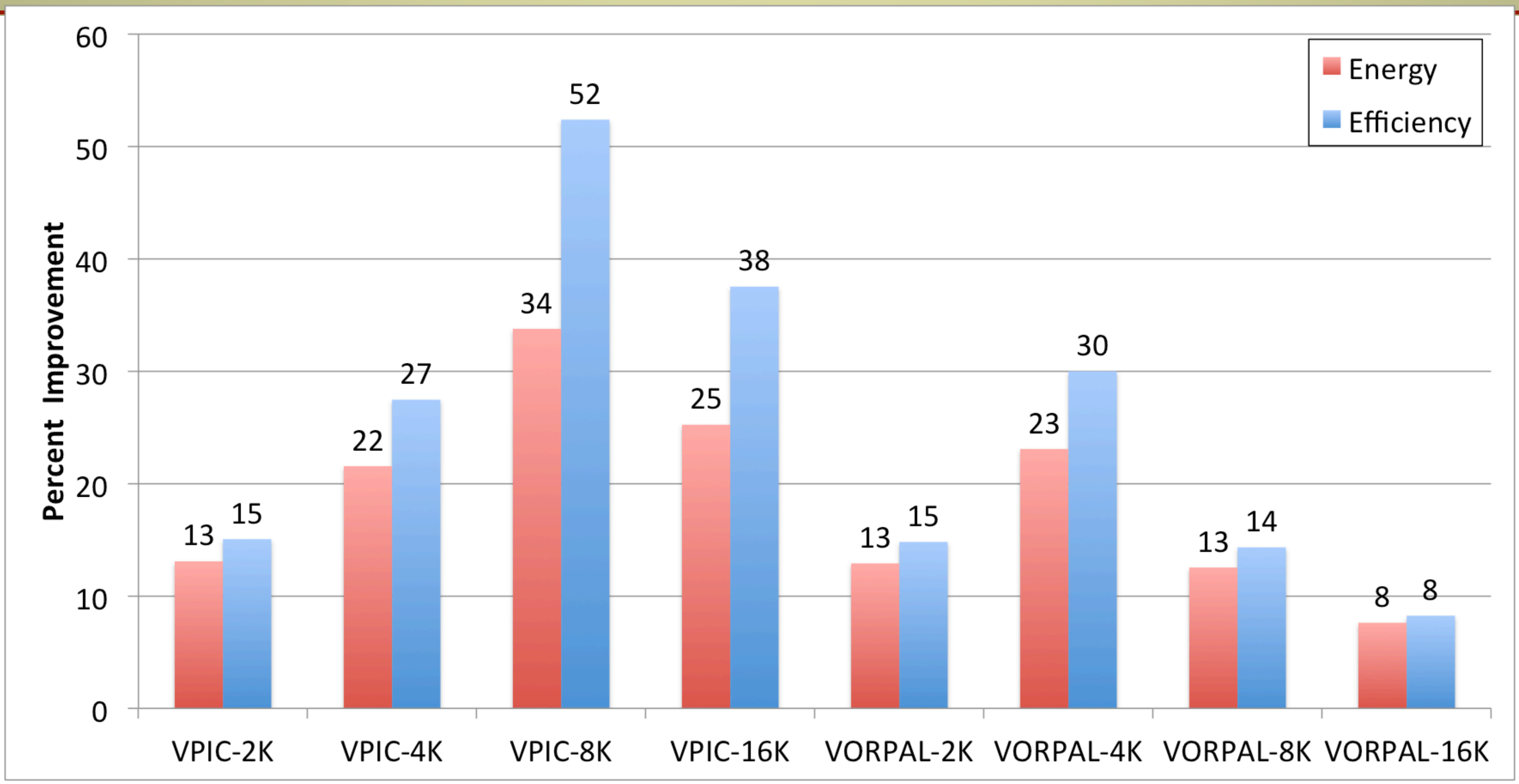


Significant I/O performance degradation at low frequencies

Weak-scaling – Energy Efficiency



Weak scaling – Energy Savings & Improvement

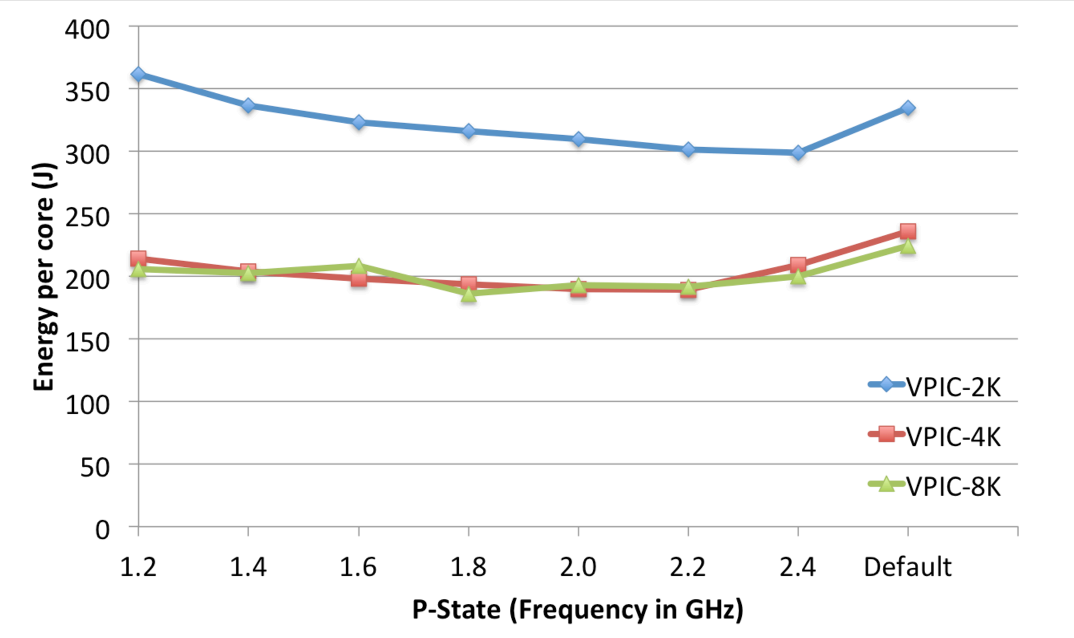


Least energy consumption to default energy

Highest energy efficiency to default energy efficiency

- 5 out of 8 @ 2.2 GHz
- Others @ 1.8 and 1.4 GHz

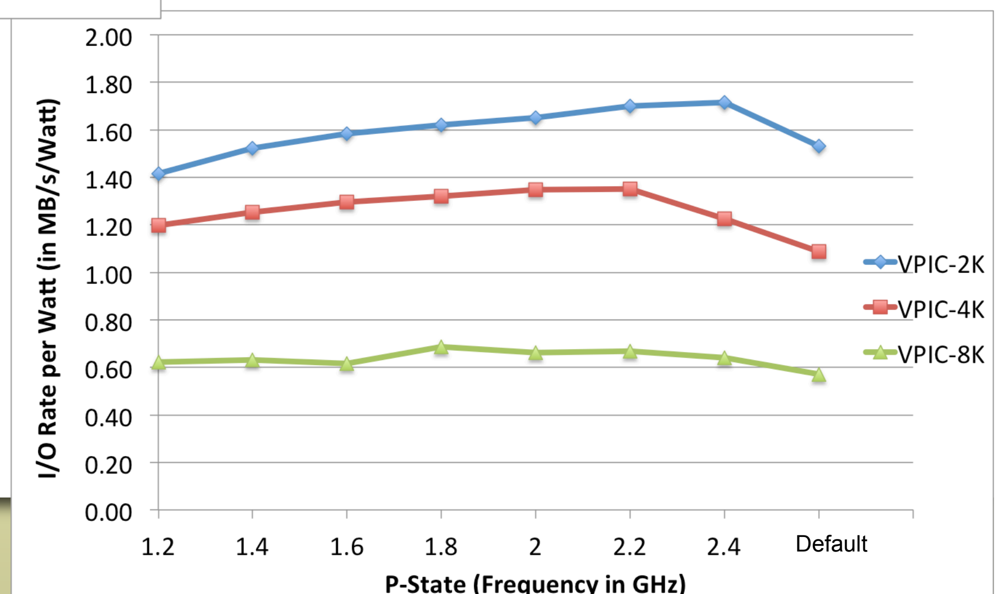
Strong Scaling – Energy and Energy Efficiency



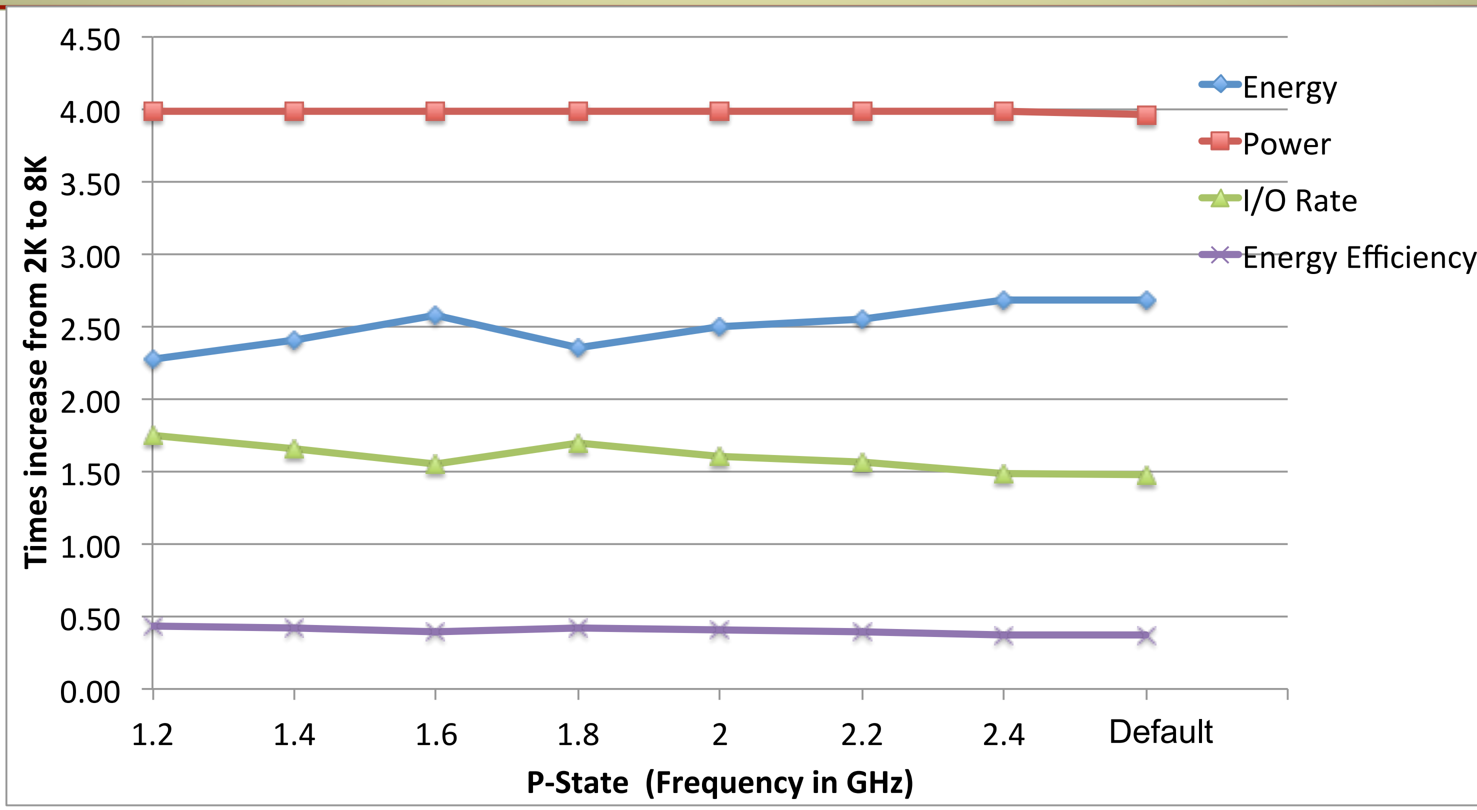
Better

Best:
2K → 2.4 GHz
4K → 2.2 GHz
8K → 1.8 GHz

Better



Strong Scaling – Trends



Power increases by 4X from 2K to 8K
Energy efficiency decreases by 50%

Observations & Unknowns

- Decreased I/O rate with frequency?
  - CPU and node activity during I/O phase
  - Using fewer cores per node or pinning fewer cores to perform I/O
  - MPI-IO in independent mode
  - I/O performance variation
  - Fine grain power state settings – Some cores at high frequency and some at lower
- I/O phase energy consumption with new memory and storage hierarchy?
  - Node level NVM
  - Burst buffers

Thanks!

Advanced Scientific Computing Research (ASCR) for funding the Power-aware Data Management project
Program Manager: Lucy Nowell
Project PI: Hank Childs