Project Zeta: an integrated simulation and analysis platform for earth system science

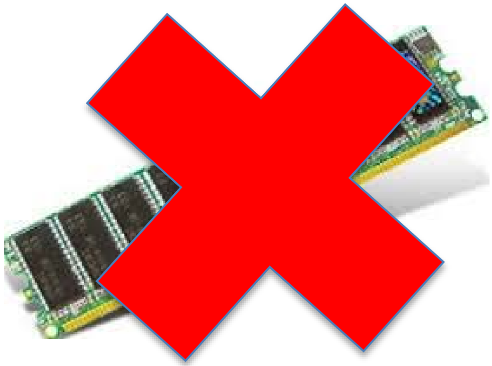
Dr. Richard Loft
Director, Technology Development
Computational and Information Systems Laboratory
National Center for Atmospheric Research

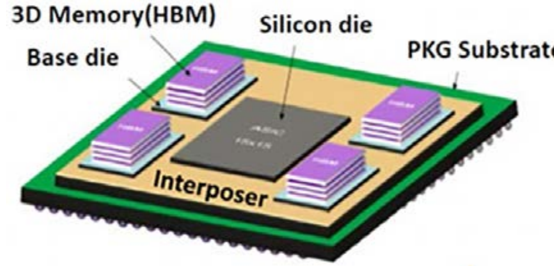
ZETA = ZEro-copy Trans-petascale Architecture

Application developer's view of exascale technology

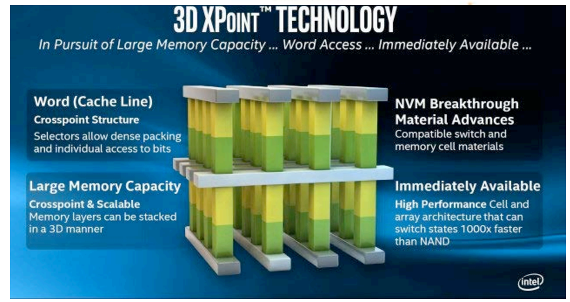
Credit: Fast and Furious 8

New technologies, faster science?

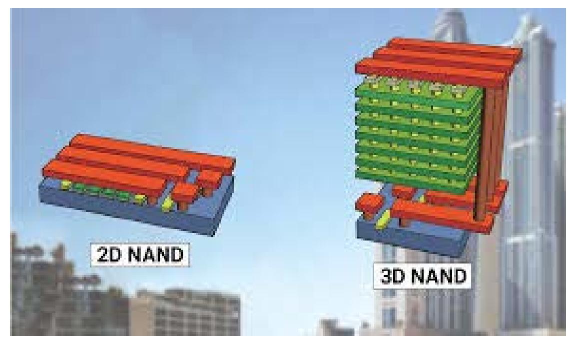




Stacked memory:
Fast, hot & small

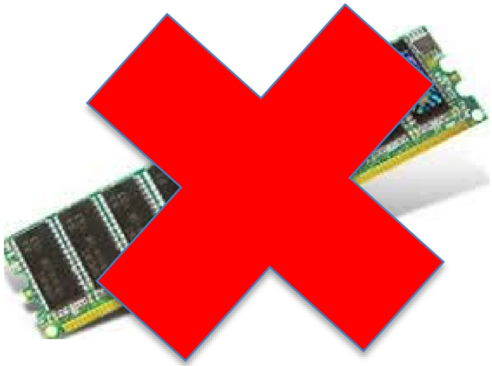


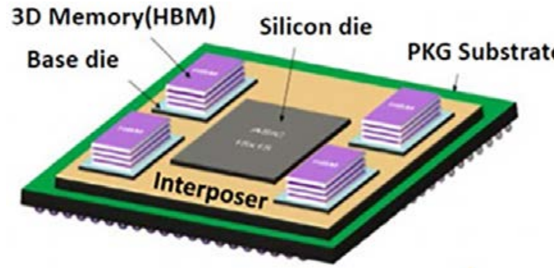
Memory-class storage



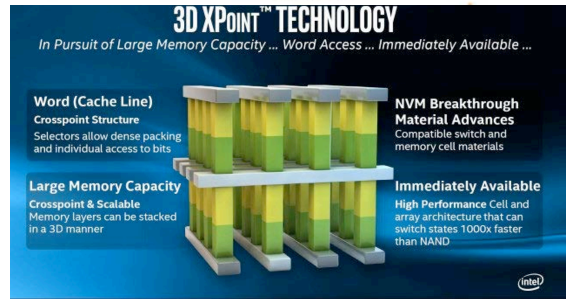
Storage-class memory

New technologies, faster science?

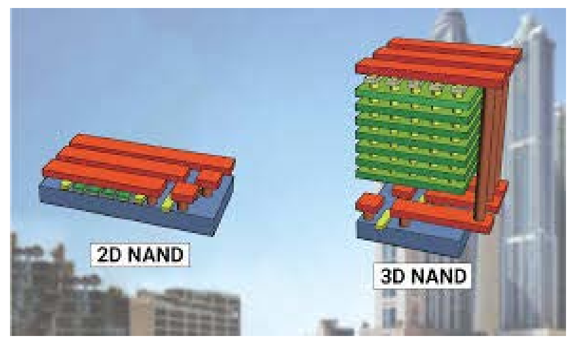




Stacked memory:
Fast, hot & small



Memory-class storage



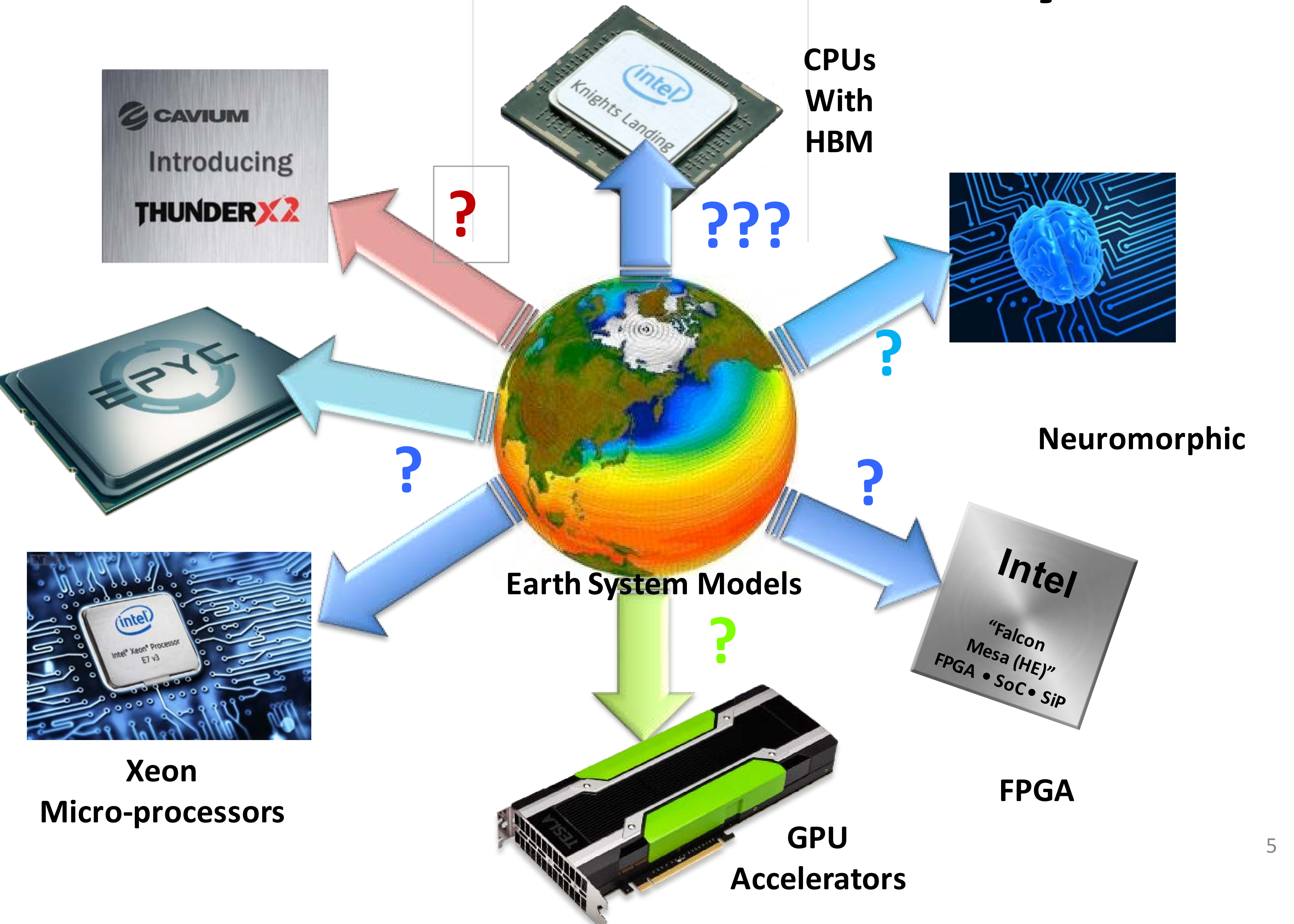
Storage-class memory

S3

ceph

Cloud-base
object store
public or private)

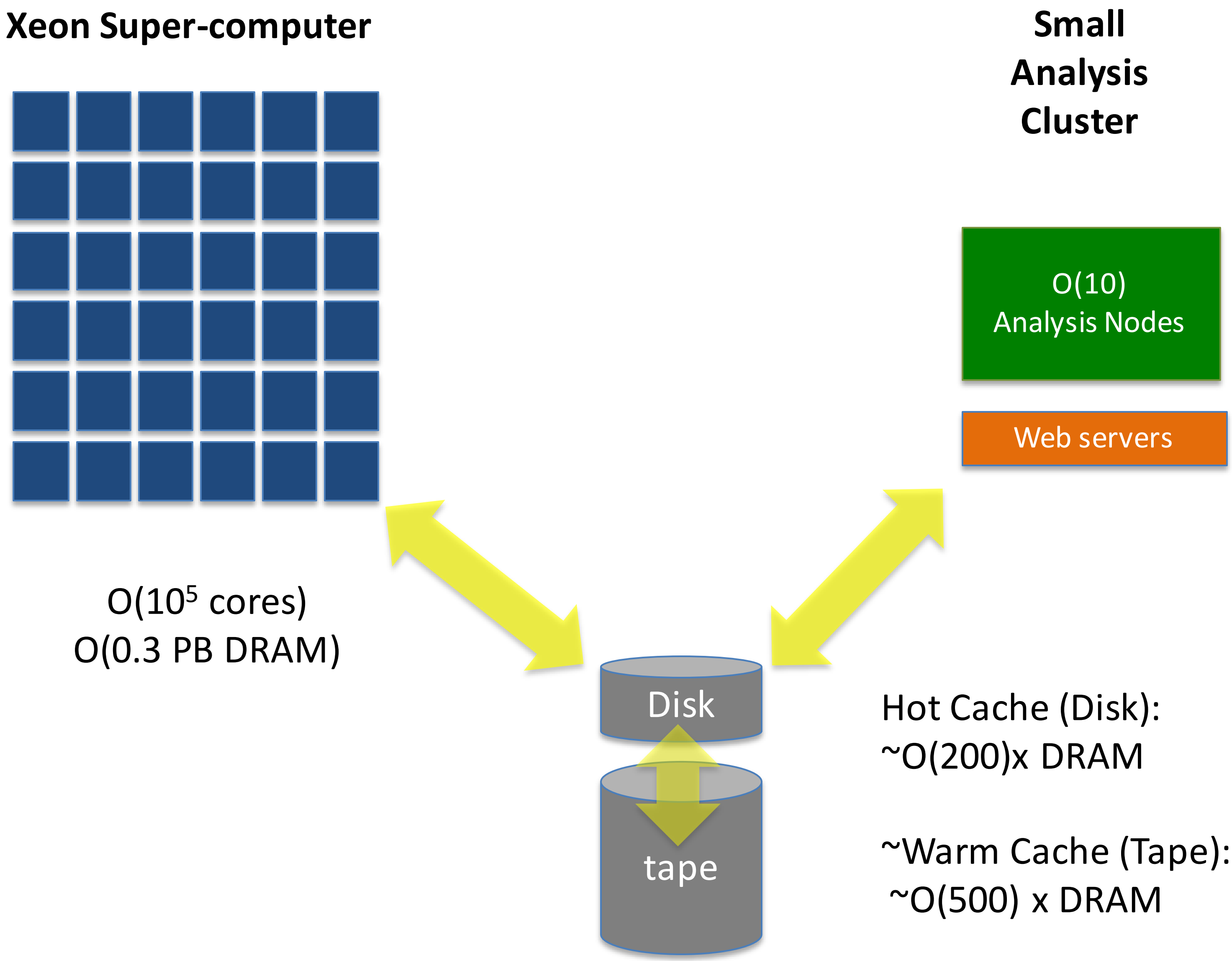
Preformance Portability?



Project Zeta Goals

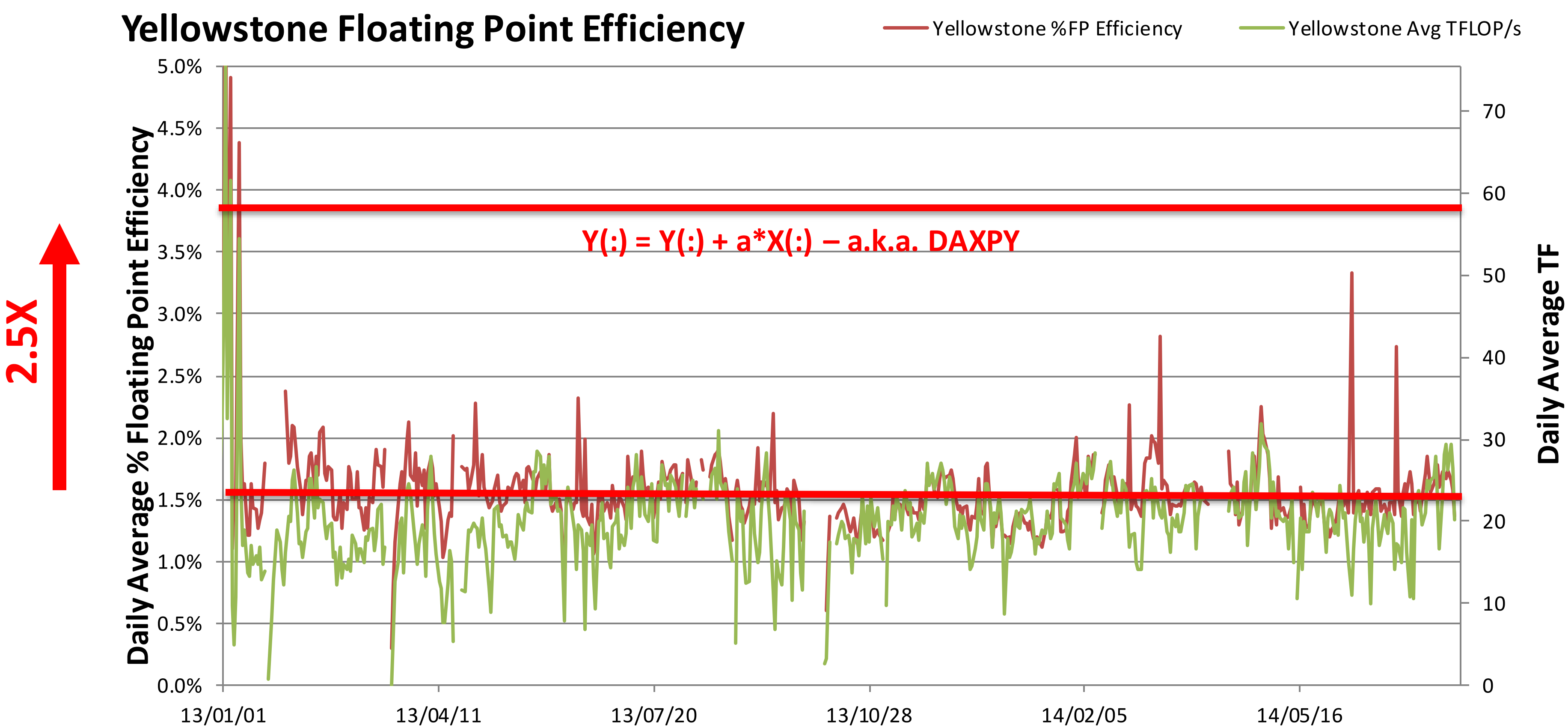
- **Focus on a design in Zeta that:**
  - Enhances the end-to-end rate of science throughput
  - Reduces costs and/or enhance reliability
- **Harness emerging technologies for Zeta like:**
  - Accelerators (GPUs)
  - New memory technologies (stacked, NV memory)
  - Machine learning techniques (DL)
- **Prepare application/workflow codes for Zeta:**
  - scalability and performance
  - Performance-portability

Existing Architecture



What's wrong with our performance?

Yellowstone: Sustained fraction of FP peak was 1.57%

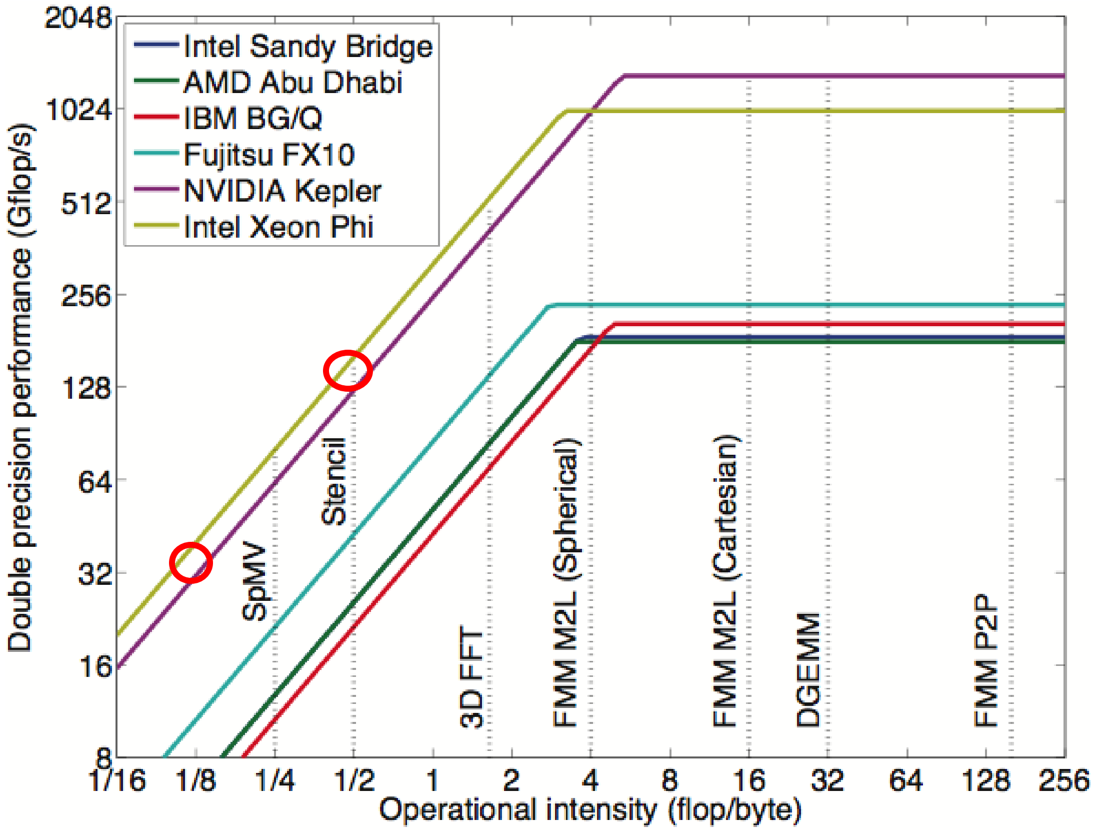


Knowing your limits:
the roofline diagram

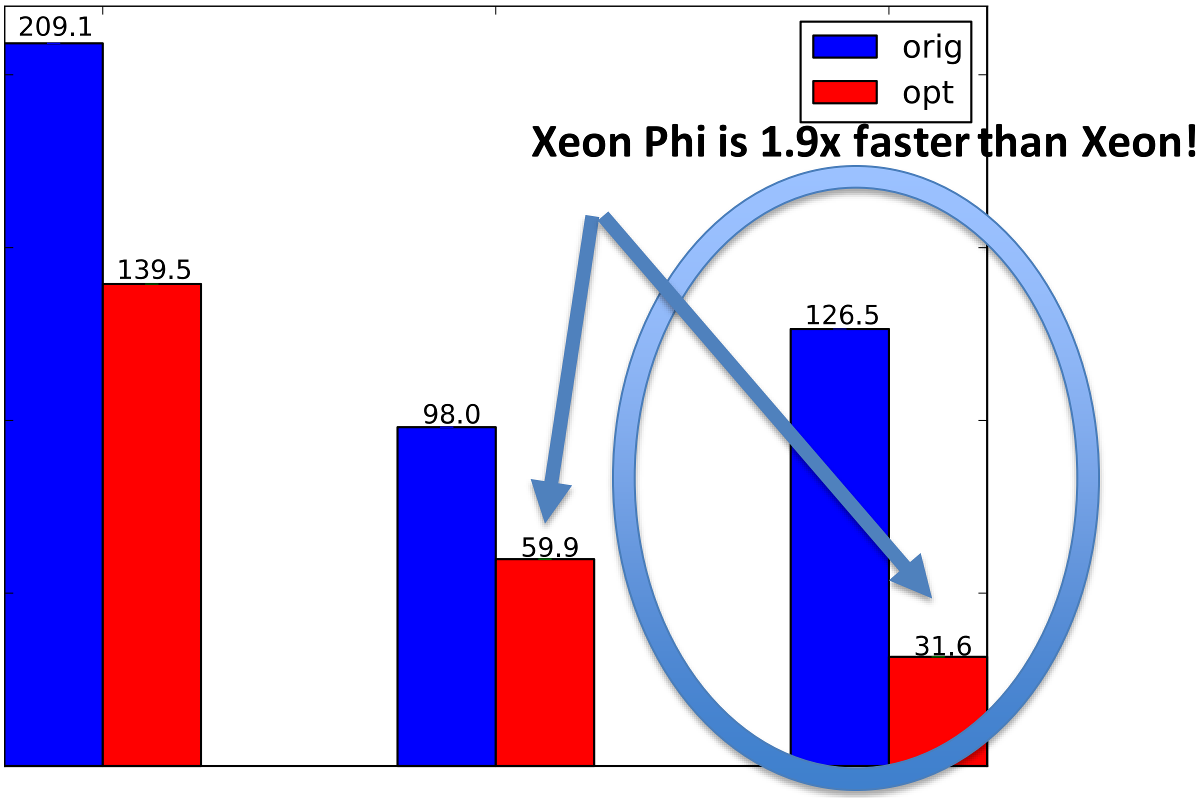
Source: Barba and Yokota, SIAM News, Volume 46, Number 6, July/August 2013

MOM6 barotropic stencil
0.125 flop/byte (DP)

RBF-FD SWE Model
0.5 flop/byte (DP)

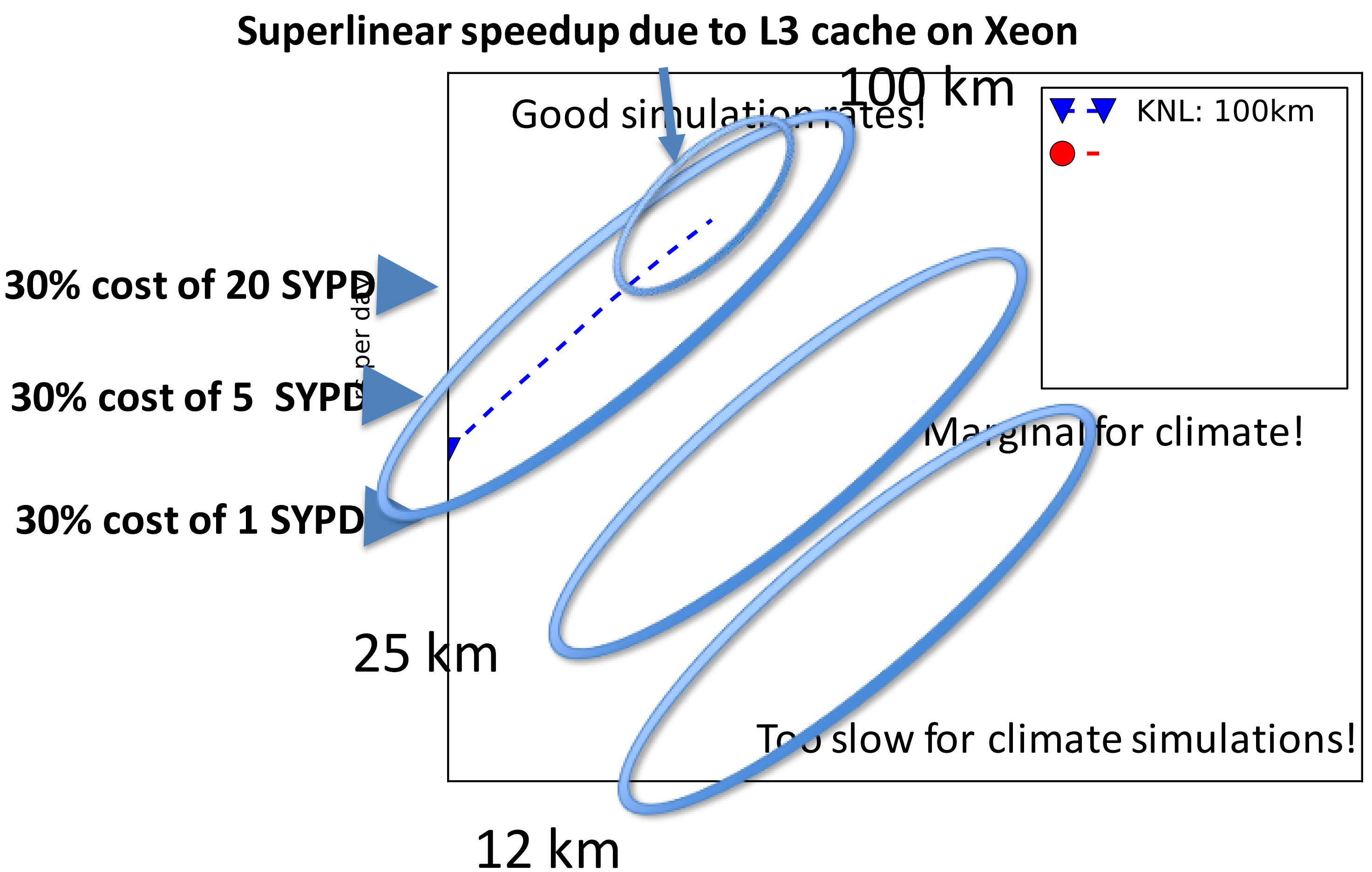


HOMME (NE=8, PLEV=70, qsize=135)



75% reduction in cost!

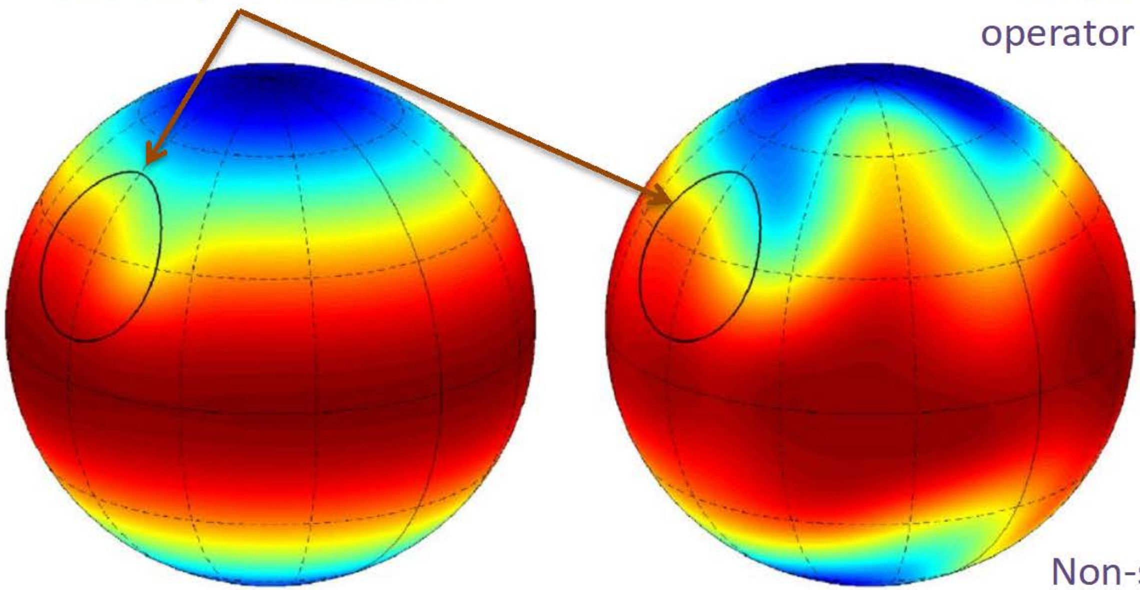
Simulation rate for HOMME on Xeon and KNL



Optimizing Stencils for different architectures

Benchmark Problem

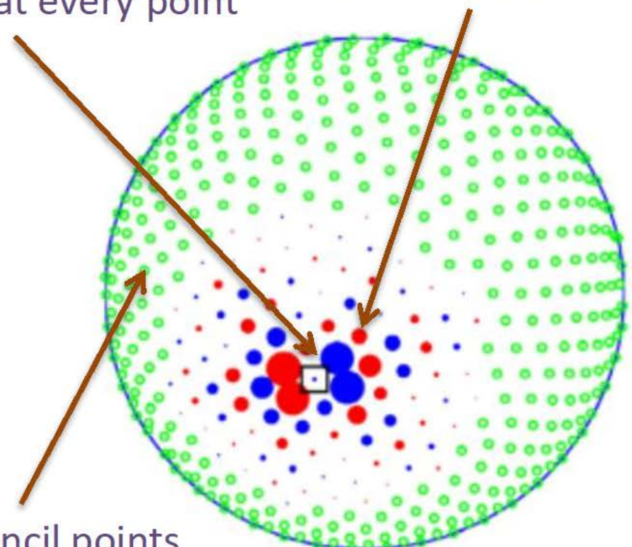
- Shallow Water Equations (SWE)
  - A set of non-linear partial differential equations (PDE)
  - Capture features of atmospheric flow around the Earth
- Radial basis function-generated finite difference (RBF-FD) methods



Day 1

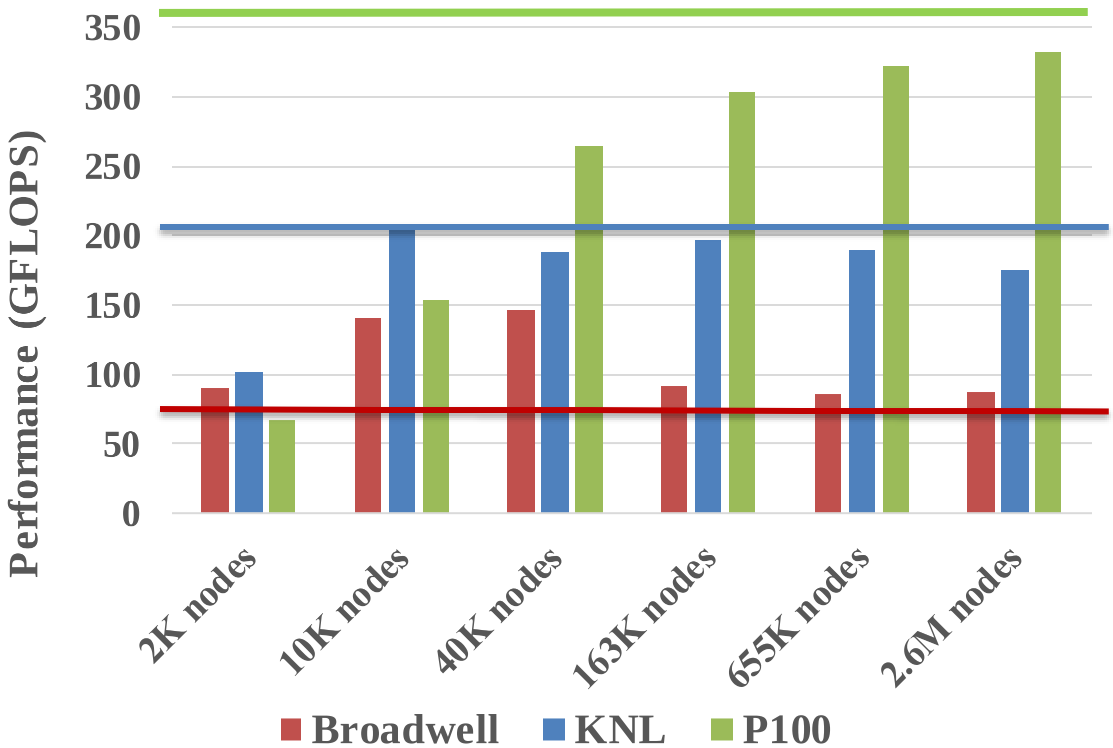
Day 15

RBF-FD solution to SWE test case “Flow over an isolated mountain” using 655,532 points [1]



An example of 75-point stencil on a sphere [1]

CISL experiences with directive-based portability: RBF-FD shallow water equations: 2D unstructured stencil



Insufficient Workload Parallelism ⟷ **Sufficient Workload Parallelism**

- **CI roofline model generally predicts performance well, even for more complicated algorithms.**
- **Xeon performance crashes to DRAM BW limit when cache size is exceeded, with some state reuse.**
- **Xeon Phi (KNL) HBM memory is less sensitive to problem size that Xeon, saturates with CI figure.**
- **NVIDIA Pascal P100 performance fits CI model GPU's require higher levels of parallelism to reach saturation.**

MPAS 5 Performance

Execution time for single timestep (in seconds)

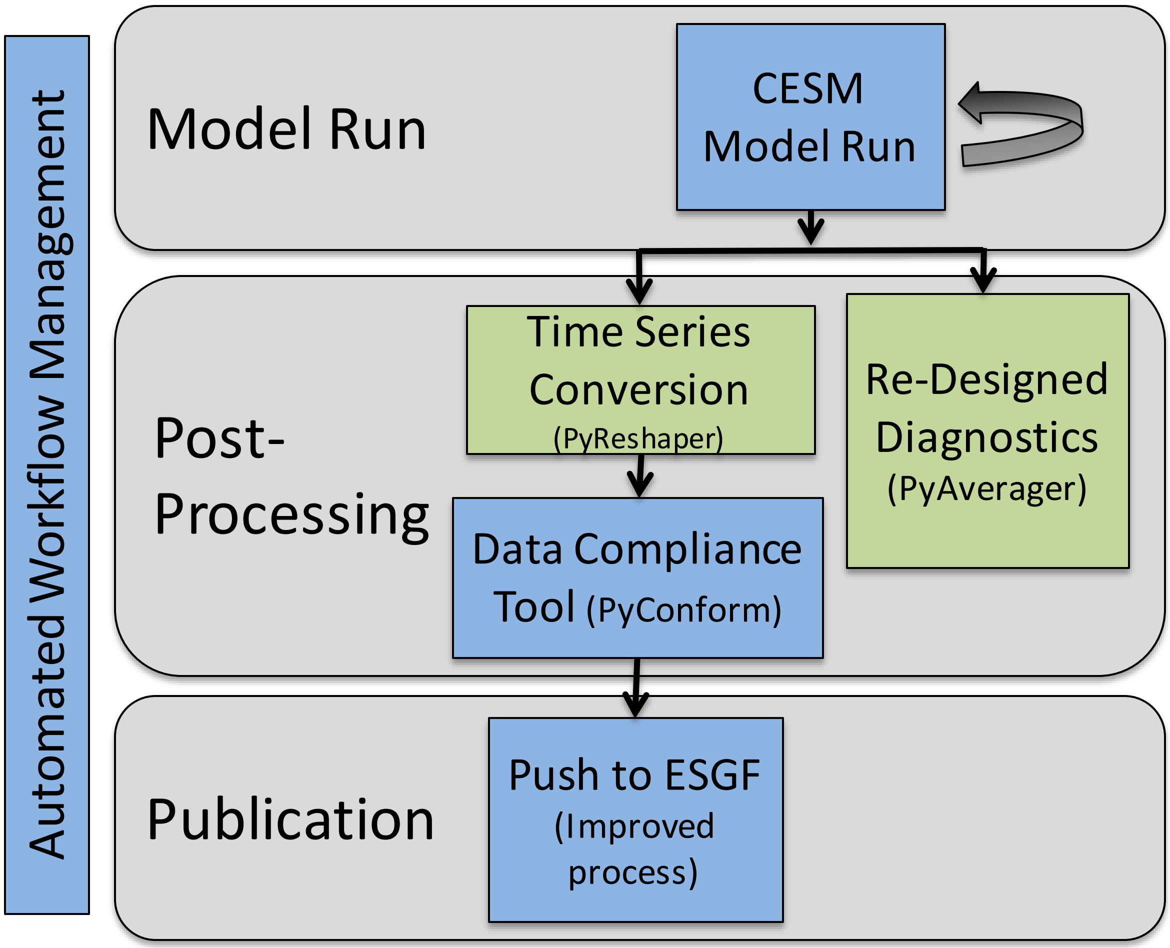
Kernels	Broadwell Node		Pascal P100		Speed Up	
	120 Km	60 Km	120 Km	60 Km	120 Km	60 Km
Integration Setup	1.21E-02	5.31E-02	1.86E-03	5.65E-03	6.51	9.40
Moist coefficients	2.08E-03	9.28E-03	1.49E-03	5.49E-03	1.40	1.69
imp_coef	4.66E-03	1.28E-02	3.20E-03	1.00E-02	1.46	1.27
dyn_tend	3.91E-03	1.41E-01	1.41E-02	4.65E-02	0.28	3.03
small_step	3.20E-02	1.44E-02	1.08E-03	3.81E-03	29.67	3.77
acoustic_step	3.70E-03	3.78E-02	4.70E-03	1.81E-02	0.79	2.09
large_step	1.03E-02	5.09E-02	2.78E-03	1.04E-02	3.71	4.90
diagnostics	1.63E-02	8.22E-02	4.53E-03	1.75E-02	3.59	4.68
Time step Loop	0.92	3.49	0.37	1.26	2.48	2.76

Code currently being upgraded to MPAS 5.2

NCAR performance portability experiences…

- Refactoring code for vectorization can yield ~**2.5-4x** performance improvements for x86 multi-/many-cores. We've been co-designing a vectorizing ifort….
- Directive-based parallelism provides portability across Xeon, Xeon-Phi and GPU. Maintaining single source feasible for many cases (RBFs & MPAS).
- OpenACC is in a sense a "domain specific language". We've been co-designing OpenACC with PGI…
- Would be nice if a std emerge (e.g. OpenMP)
- Portability across 3 architectures is all great but…

CESM/CMIP6 Workflow



NCAR Analytics Accomplishments: The Low Hanging Fruit

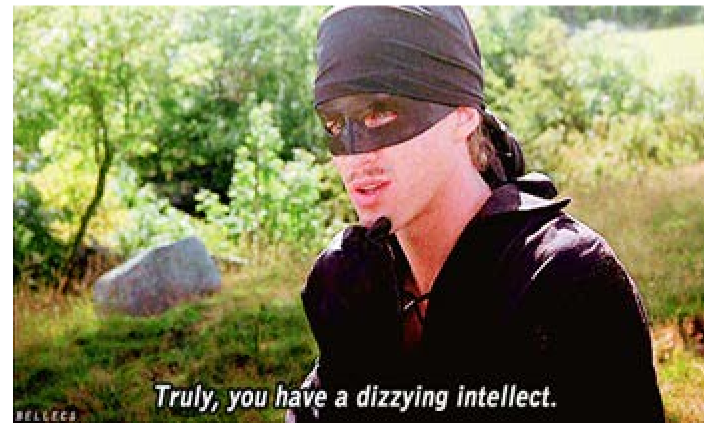
- Parallel tools: **PyReshaper, PyAverager, PyConform**
- Parallelizing **PyReshaper** yielded **~6.5x** on Edison
- **NAND-based** tests
  - Py{*} analytics **2.5-6x**
  - subsetting (RDA) **20x**
- Automating workflows (Cycl) saved **O(3x)**
- **5x** storage volume savings through lossy data compression (discussed yesterday).

Unsupervised Learning: Generative Adversarial Networks

Unsupervised method of learning complex feature representations from data
Requires 2 deep neural networks

Discriminator: determines which samples are from the training set and which are not

Generator: Creates synthetic examples similar to training data to fool discriminator



Both networks have a "battle of wits" either to the death or until the discriminator is fooled often enough

Advantages

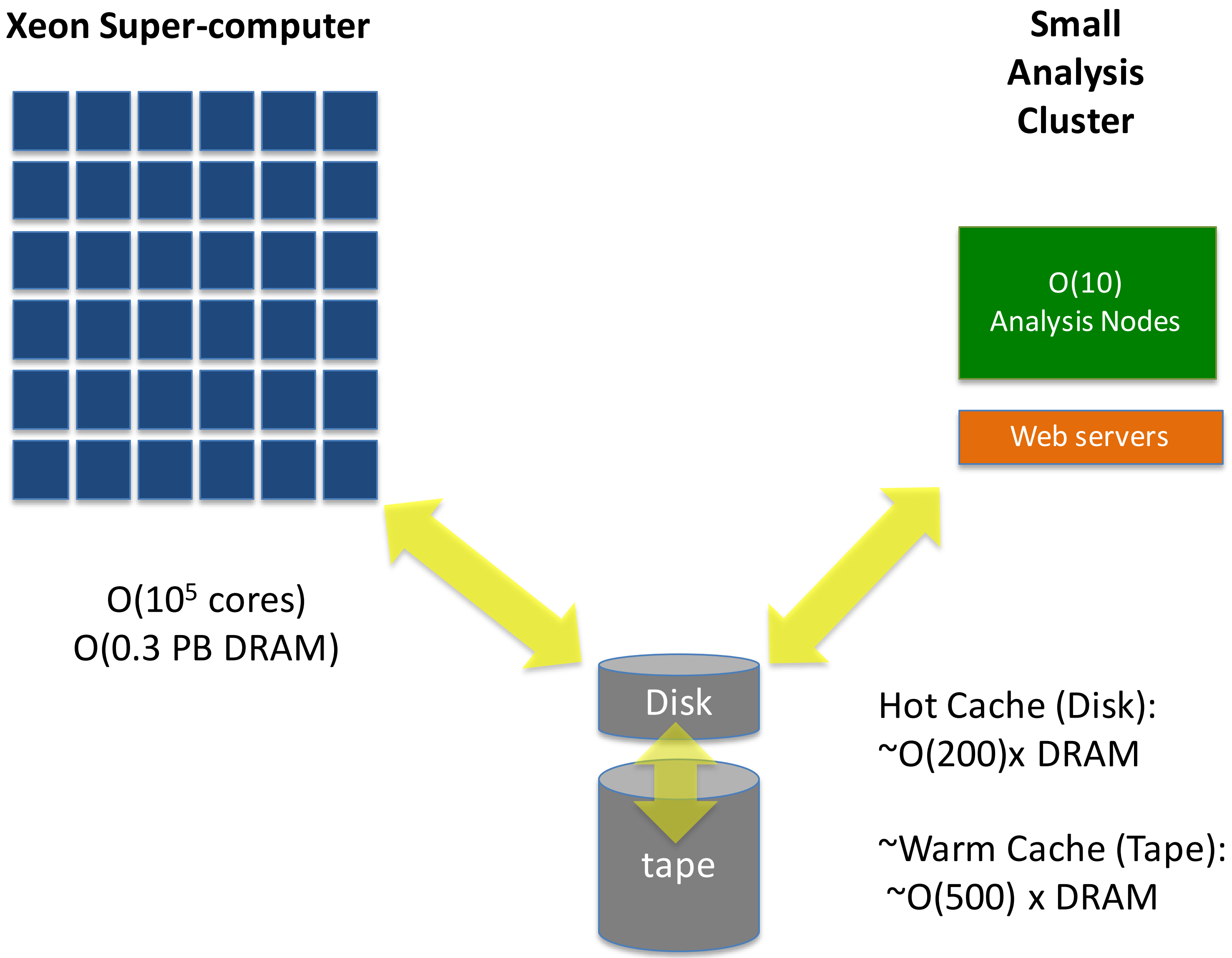
- **Unsupervised pre-training: learn features without needing a large labeled dataset**
- **Dimensionality reduction: reduce image to smaller vector**
- **Learns sharper, more detailed features than auto-encoder models**
- **Do not need to specify a complex loss function**

Credit: Princess Bride

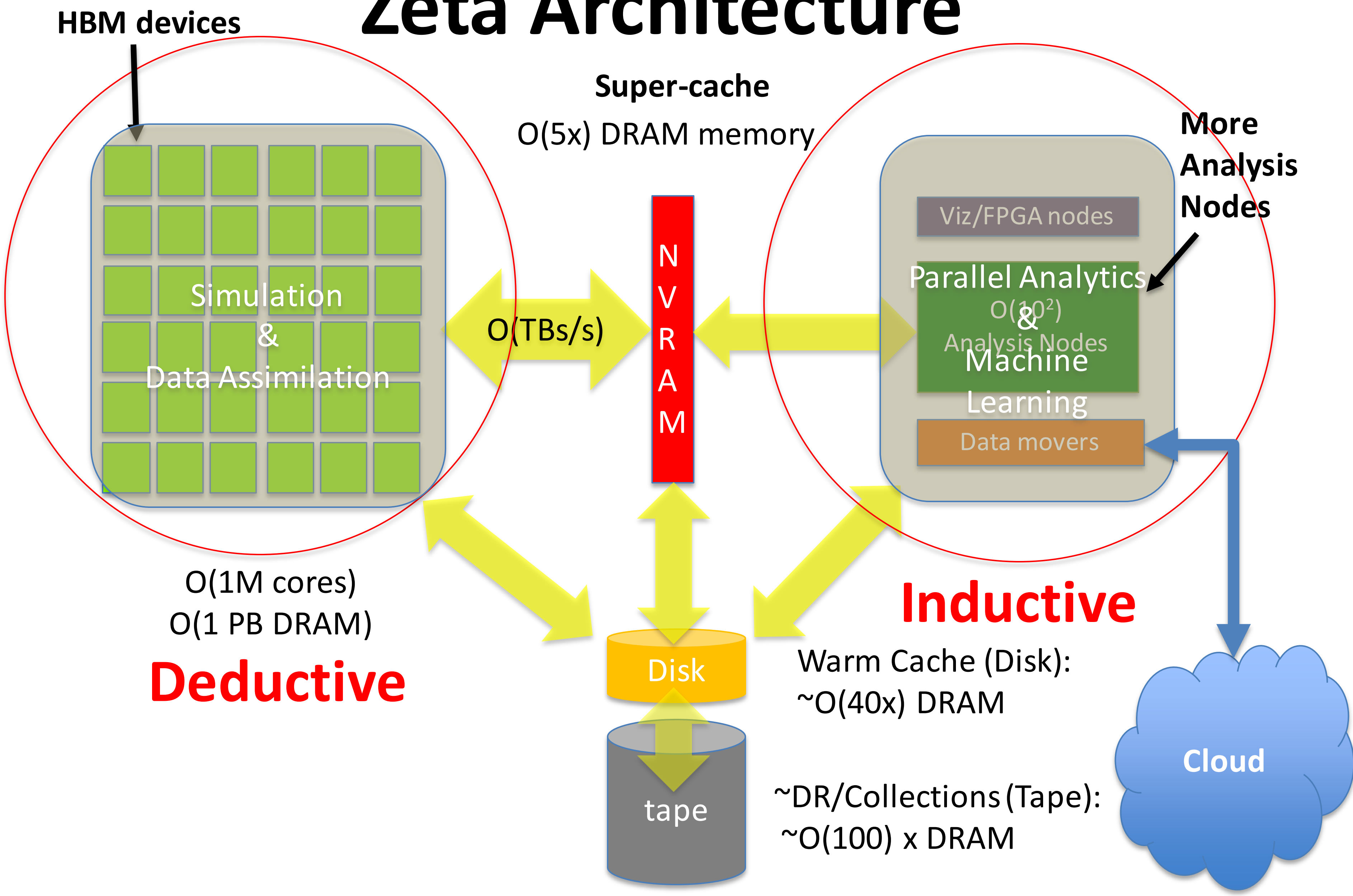
Pros and cons of building DL emulators

- **Pros**
  - Drafts behind DL-driven technology
  - May be less (80x?) computationally intensive
  - Deep Learning leverages frameworks.
  - Less code to develop (code is in the weights and the network design)
- **Cons**
  - Potential loss of understanding of the physical basis of results.
  - Over-fitting, curse of dimensionality, etc. Kind of an art.
  - Not clear how conservation laws/constraints are preserved in DL systems.

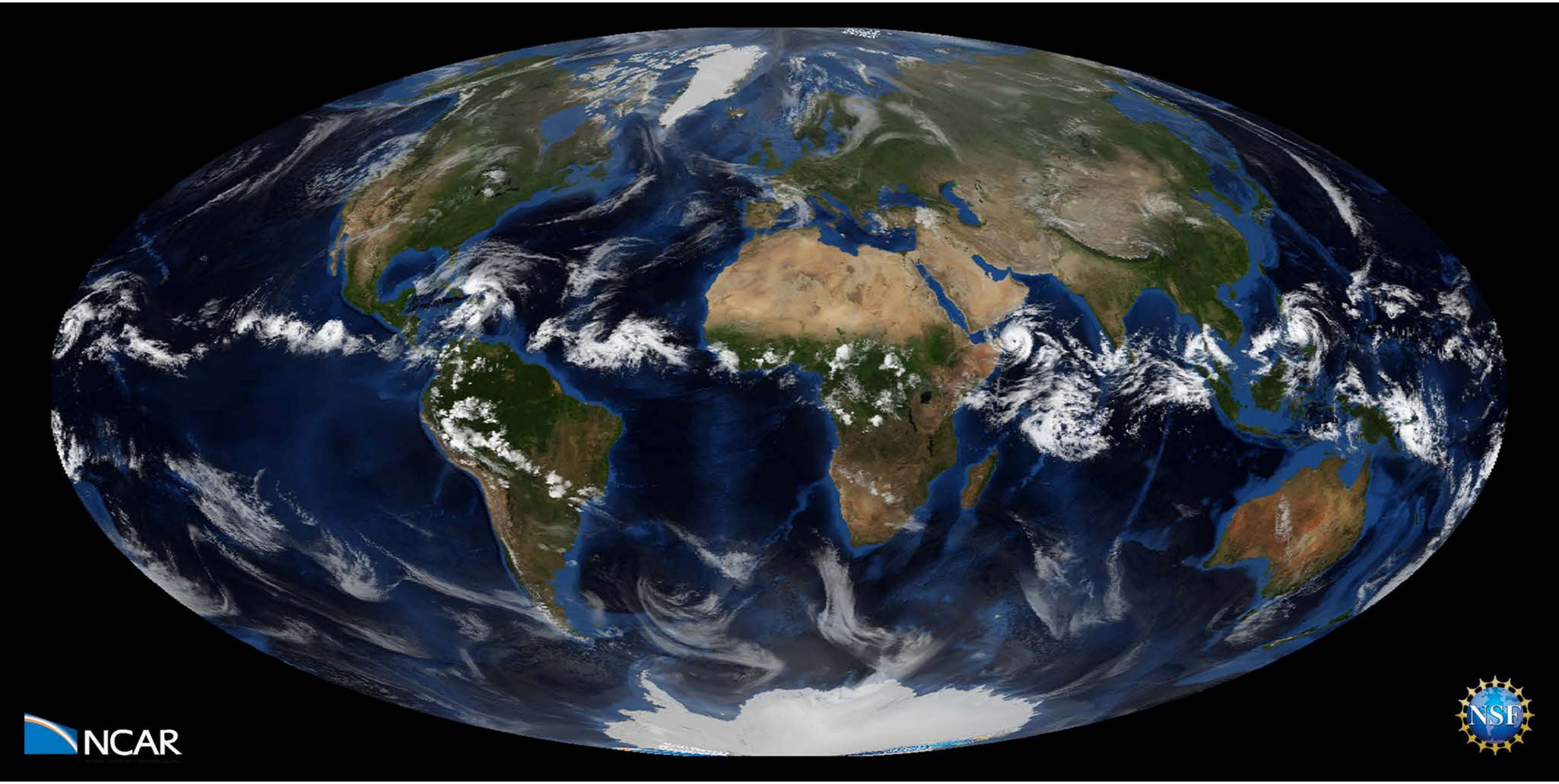
Existing Architecture



Zeta Architecture



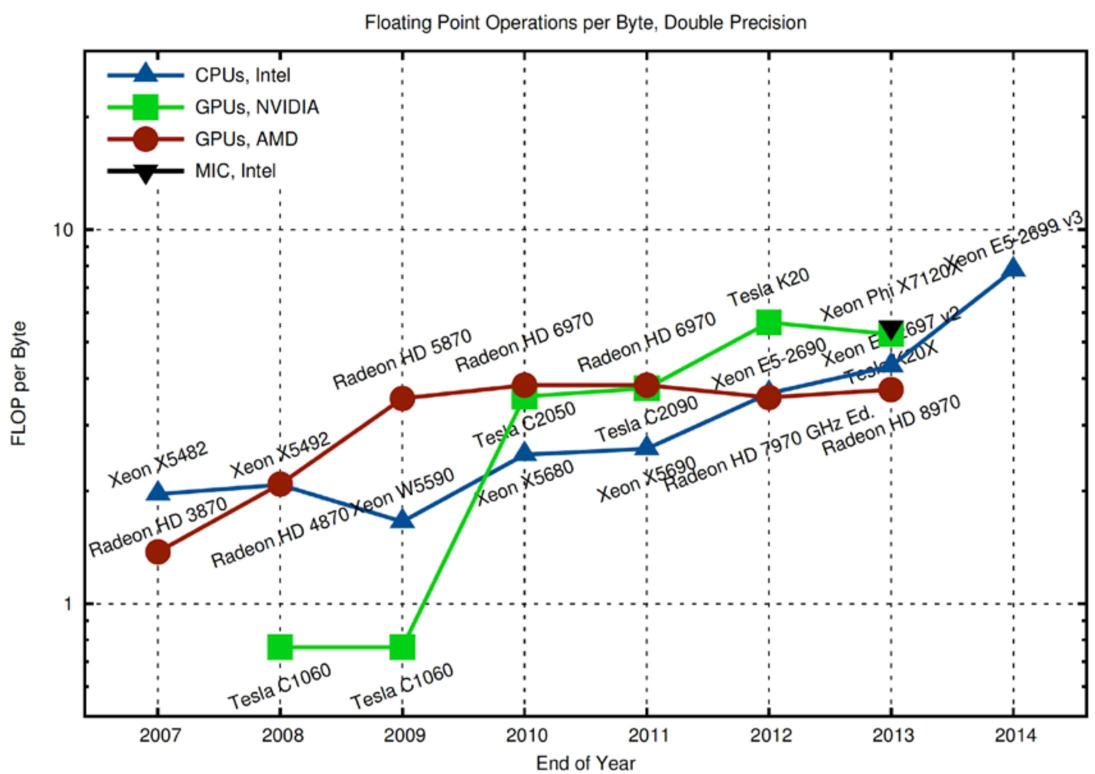
Thanks!



Current supercomputers struggle on HPCG relative to HP Linpack:

Site	Computer	Cores	HPL Rmax (Pflops)	HPL Rank	HPCG (Pflops)	HPCG/ HPL	% of Peak
NSCC / Guangzhou	Tianhe-2 NUDT, Xeon 12C 2.2GHz + Intel Xeon Phi 57C + Custom	3,120,000	33.9	1	.632	1.8%	1.1%
RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx 8C + Custom	705,024	10.5	4	.461	4.4%	4.1%
DOE/OS Oak Ridge Nat Lab	Titan, Cray XK7 AMD 16C + Nvidia Kepler GPU 14C + Custom	560,640	17.6	2	.322	1.8%	1.2%
DOE/OS Argonne Nat Lab	Mira BlueGene/Q, Power BQC 16C 1.60GHz + Custom	786,432	8.59	5	.167	1.9%	1.7%
Swiss CSCS	Piz Daint, Cray XC30, Xeon 8C + Nvidia Kepler 14C + Custom	115,984	6.27	6	.105	1.7%	1.3%
Leibniz Rechenzentrum	SuperMUC, Intel 8C + IB	147,456	2.90	14	.0833	2.9%	2.6%
DOE/OS LBNL	Edison, Cray XC30, Xeon, 12c, 2,4GHz + Custom	133,824	1.65	24	.0786	4.8%	3.1%
GSIC Center TiTech	Tsubame 2.5 Xeon 6C, 2.93GHz + Nvidia K20x + IB	76,032	2.78	15	.073	2.6%	1.3%
Max-Planck	iDataPlex Xeon 10C, 2.8GHz + IB	65,320	1.28	34	.061	4.8%	4.2%
CEA/TGCC-GENCI	Curie tine nodes Bullx B510 Intel Xeon 8C 2.7 GHz + IB	77,184	1.36	33	.051	3.8%	3.1%
Exploration and Production Eni S.p.A.	HPC2, Intel Xeon 10C 2.8 GHz + Nvidia Kepler 14C + IB	62,640	3.00	12	.0489	1.6%	1.2%

Processor flops/byte: trending upwards



[c/o Karl Rupp]

Energy usage for HOMME on Xeon and Xeon Phi @ 100 km