Living at the Top of the Top500: Myopia from Being at the Bleeding Edge

Bronson Messer

Oak Ridge Leadership Computing Facility
&
Theoretical Astrophysics Group
Oak Ridge National Laboratory

Department of Physics & Astronomy
University of Tennessee

Outline

- Statements made without proof
- OLCF's Center for Accelerated Application Readiness
- Speculations on task-based approaches for multiphysics applications in astrophysics (e.g. blowing up stars)

Riffing on Hank's fable...

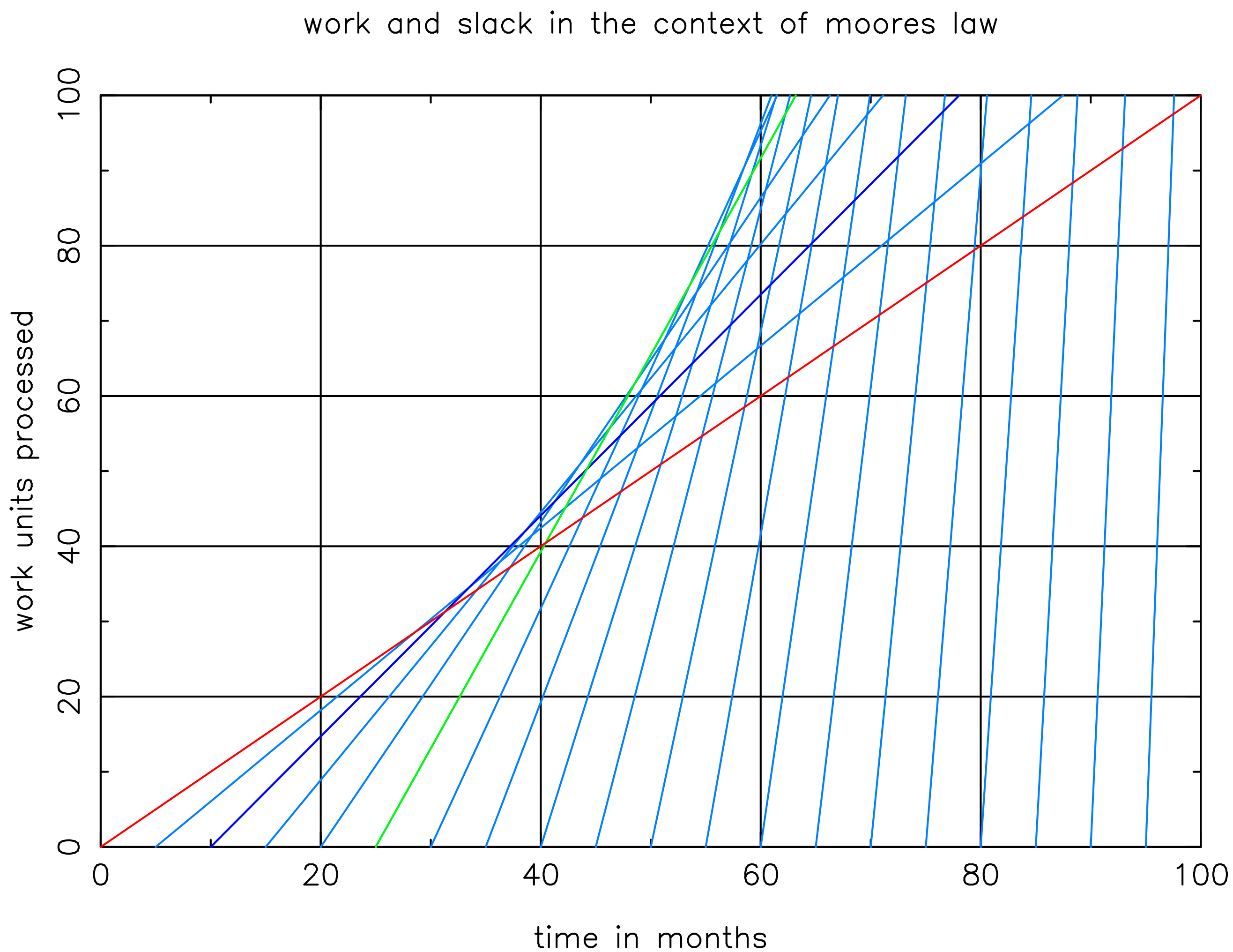
The Effects of Moore's Law and Slacking [1] on Large Computations

astro-ph/9912202

Chris Gottbrath, Jeremy Bailin, Casey Meakin, Todd Thompson, J.J. Charfman
Steward Observatory, University of Arizona

Abstract

We show that, in the context of Moore's Law, overall productivity can be increased for large enough computations by 'slacking' or waiting for some period of time before purchasing a computer and beginning the calculation.



[1] This paper took 2 days to write

OLCF
OAK RIDGE National Laboratory

Some realities

- **The future is now: if you go from franklin to hopper at the same size, you lose.**

NERSC-6 Grace "Hopper"

Cray XE6

Performance
- 1.2 PF Peak
- 1.05 PF HPL (#5)

Processor
- AMD MagnyCours
- 2.1 GHz 12-core
- 8.4 GFLOPs/core
- 24 cores/node
- 32-64 GB DDR3-1333 per node

System
- Gemini Interconnect (3D torus)
- 6392 nodes
- 153,408 total cores

I/O
- 2PB disk space
- 70GB/s peak I/O Bandwidth

NERSC Franklin - Cray XT4

- 38,288 compute cores
- 9,572 compute nodes
- One quad-core AMD 2.3 GHz Opteron processors (Budapest) per node
- 4 processor cores per node
- 8 GB of memory per node
- 78 TB of aggregate memory
- 1.8 GB memory / core for applications
- /scratch disk default quota of 750 GB

- Light-weight Cray Linux operating system
- No runtime dynamic, shared-object libs
- PGI, Cray, Pathscale, GNU compilers

Some realities

- **If you use primarily IBM platforms, you have a bit longer.**
  - scp+make on Blue Waters will likely give you a speedup.
  - BG/P --> BG/Q brings an increased clock, and you probably aren't engaging the Double Hummer now anyway.

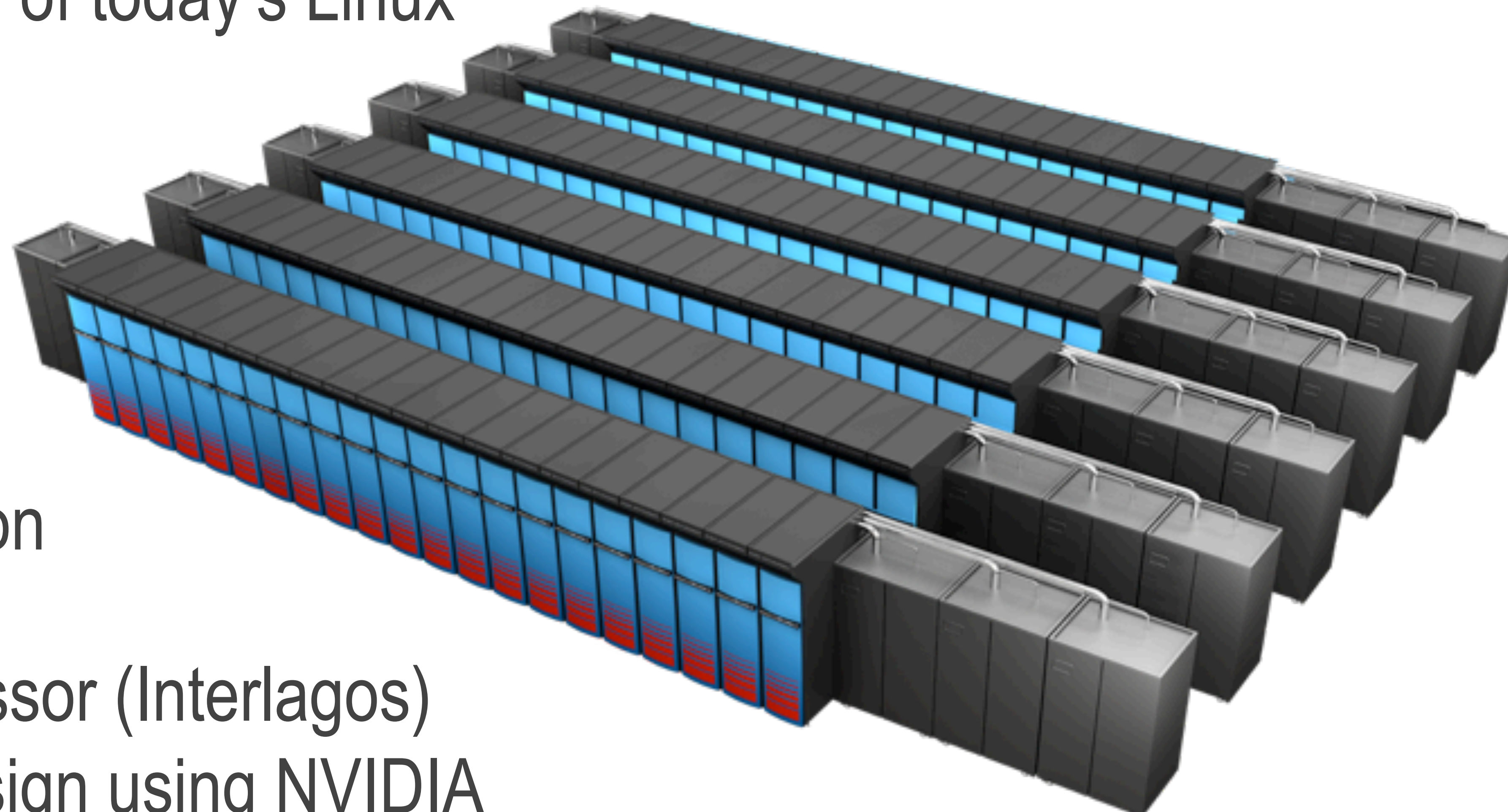
Some realities

- **It doesn't matter if you are gonna use GPU-based machines or not**
  - GPUs [CUDA, OpenCL, directives]
  - FPUs on Power [xlf, etc.]
  - Cell [SPE]
  - SSE/AVX; MIC (Knights Ferry, Knights Corner)[?]

- **Exposing the maximum amount of node-level parallelism and increasing data locality are the only way to get performance from any of these things**

ORNL's "Titan" System Goals

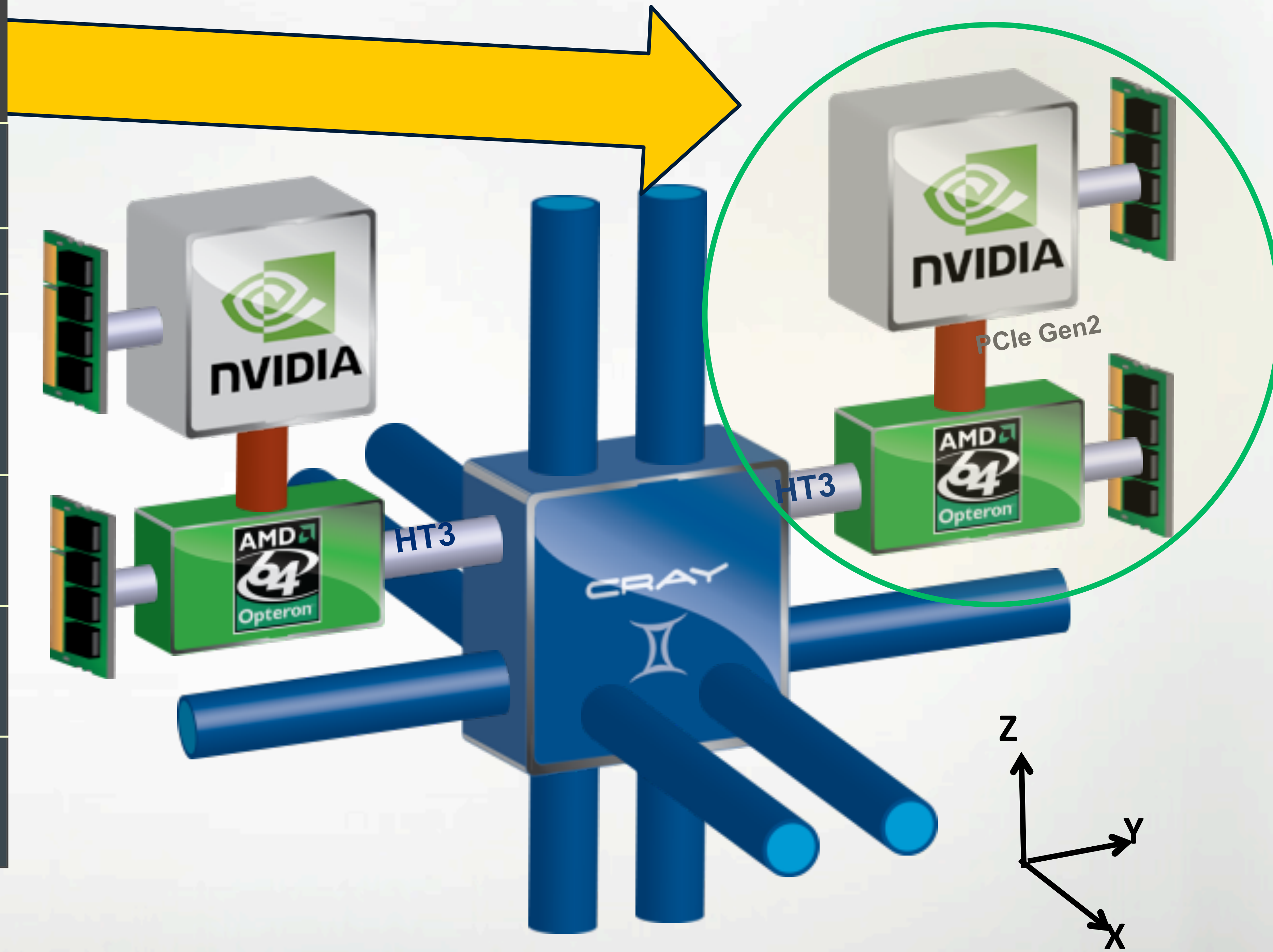
- Similar number of cabinets, cabinet design, and cooling as Jaguar
- Operating system upgrade of today's Linux operating system
- Gemini interconnect
  - 3-D Torus
  - Globally addressable memory
  - Advanced synchronization features



- AMD Opteron 6200 processor (Interlagos)
- New accelerated node design using NVIDIA multi-core accelerators
- 10-20 PF peak performance
  - Performance based on available funds
- Larger memory - more than 2x more memory per node than Jaguar

Cray XK6 Compute Node

XK6 Compute Node Characteristics
AMD Opteron 6200 Interlagos 16 core processor
Tesla X2090 @ 665 GF
Host memory 16 or 32GB 1600 MHz DDR3
Tesla X090 memory 6GB GDDR5 capacity
Gemini high speed Interconnect
Upgradeable to NVIDIA's Kepler many-core processor



Slide courtesy of Cray, Inc.

OLCF-3 Applications Requirements developed by surveying science community

OLCF Application Requirements Document

- Elicited, analyzed, and validated using a new comprehensive requirements questionnaire
- Project overview, science motivation and impact, application models, algorithms, parallelization strategy, software, development process, SQA, V&V, usage workflow, performance
- Results, analysis, and conclusions documented in 2009 OLCF application requirements document

Science Driver Survey

- Developed in consultation with 50+ leading scientists in many domains
- Key questions
  - What are the science goals and does OLCF-3 enable them?
  - What might the impact be if the improved science result occurs?
  - What does it matter if this result is delivered in the 2012 timeframe?

Science Driver Survey

- Science driver
  - What science will be pursued on this system and how is it different (in fidelity/quality/predictability and/or productivity/throughput) from the current system
- Science impact
  - What might the impact be if this improved science result occurs? Who cares, and why?
- Science timeliness
  - If this result is delivered in the 2010 timeframe, what does it matter as opposed to coming 5 years later (or never at all)? What other programs agencies, stakeholders, and/or facilities are dependent up on the timely delivery of this result, and why?

OLCF-3 Applications Analyzed

Science outcomes were elicited from a broad range of applications

Application area	Application codes	Science target
Astrophysics	Chimera, GenASiS	• Core-collapse supernovae simulation; validation against observations of neutrino signatures, gravitational waves, and photon spectra
	MPA-FT, MAESTRO	• Full-star type Ia supernovae simulations of thermonuclear runaway with realistic subgrid models
Bioenergy	LAMMPS, GROMACS	• Cellulosic ethanol: dynamics of microbial enzyme action on biomass
Biology	LAMMPS	• Systems biology • Genomic structure
Chemistry	CP2K, CPMD	• Interfacial chemistry
	GAMESS	• Atmospheric aerosol chemistry • Fuels from lignocellulosic materials
Combustion	S3D	• Combustion flame front stability and propagation in power and propulsion engines
	RAPTOR	• Internal combustion design in power and propulsion engines: bridge the gap between device- and lab-scale combustion
Energy Storage	MADNESS	• Electrochemical processes at the interfaces; ionic diffusion during charge-discharge cycles

OLCF-3 Applications Analyzed

Science outcomes were elicited from a broad range of applications

Application area	Application codes	Science target
Fusion	GTC	• Energetic particle turbulence and transport in ITER
	GTS	• Electron dynamics and magnetic perturbation (finite-beta) effects in a global code environment for realistic tokamak transport • Improved understanding of confinement physics in tokamak experiments • Address issues such as the formations of plasma critical gradients and transport barriers
	XGC1	• First-principles gyrokinetic particle simulation of multiscale electromagnetic turbulence in whole-volume ITER plasmas with realistic diverted geometry
	AORSA, CQL3D	• Tokamak plasma heating and control
	FSP	• MHD scaling to realistic Reynolds numbers • Global gyrokinetic studies of core turbulence encompassing local & nonlocal phenomena and electromagnetic electron dynamics
	GYRO, TGYRO	• Predictive simulations of transport iterated to bring the plasma into steady-state power balance; radial transport balances power input
Geoscience	PFLOTRAN	• Stability and viability of large-scale CO_2 sequestration • Predictive contaminant ground water transport

OLCF-3 Applications Analyzed

Science outcomes were elicited from a broad range of applications

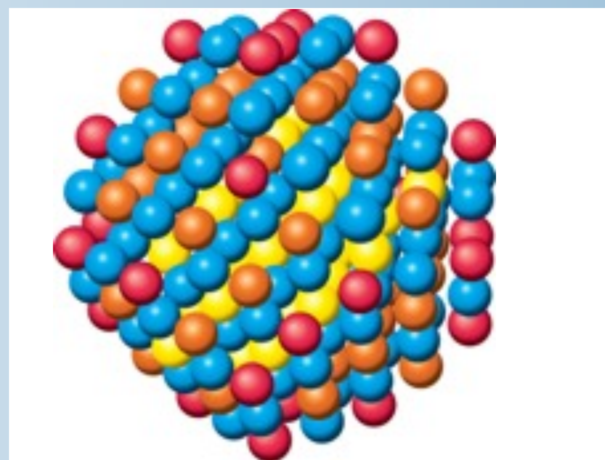
Application area	Application codes	Science target
Nanoscience	OMEN	• Electron-lattice interactions and energy loss in full nanoscale transistors
	LS3DF	• Full device simulation of a nanostructure solar cell
	DCA++	• Magnetic/superconducting phase diagrams including effects of disorder • Effect of impurity configurations on pairing and the high-T superconducting gap • High-T superconducting transition temperature materials dependence in cuprates
	WL-LSMS	• To what extent do thermodynamics and kinetics of magnetic transition and chemical reactions differ between nano and bulk? • What is the role of material disorder, statistics, and fluctuations in nanoscale materials and
Nuclear energy	Denovo	• Predicting, with UQ, the behavior of existing and novel nuclear fuels and reactors in transient and nominal operation
	UNIQ	• Reduce uncertainties and biases in reactor design calculations by replacing existing multi-level homogenization techniques with more direct solution methods
Nuclear Physics	NUCCOR MFDn	• Limits of nuclear stability, static and transport properties of nucleonic matter • Predict half-lives, mass and kinetic energy distribution of fission fragments and fission cross sections
QCD	MILC,Chroma	• Achieving high precision in determining the fundamental parameters of the Standard Model (masses and mixing strengths of quarks)
Turbulence	DNS	• Stratified and unstratified turbulent mixing at simultaneous high Reynolds and Schmidt numbers
	Hybrid	• Nonlinear turbulence phenomena in multi-physics settings

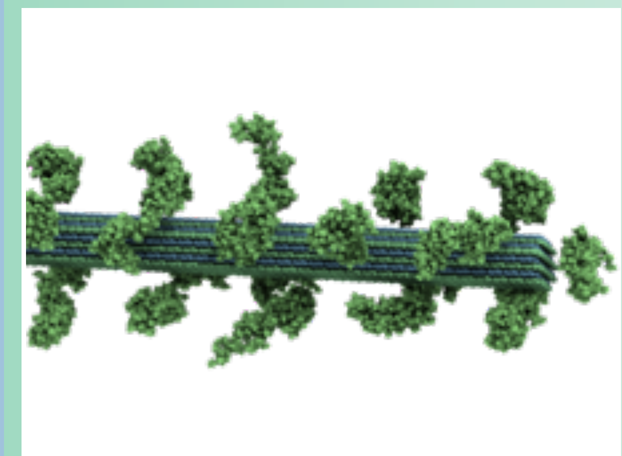
Evaluation Criteria for Selection of Six Representative Applications

Task	Description
Science	• Science results, impact, timeliness • Alignment with DOE and U.S. science mission (CD-0) • Broad coverage of science domains
Implementation (models, algorithms, software)	• Broad coverage of relevant programming models, environment, languages, implementations • Broad coverage of relevant algorithms and data structures (motifs) • Broad coverage of scientific library requirements
User community (current and anticipated)	• Broad institutional and developer/user involvement • Good representation of current and anticipated INCITE workload
Preparation for steady state ("INCITE ready") operations	• Mix of low ("straightforward") and high ("hard") risk porting and readiness requirements • Availability of OLCF liaison with adequate skills/experience match to application • Availability of key code development personnel to engage in and guide readiness activities

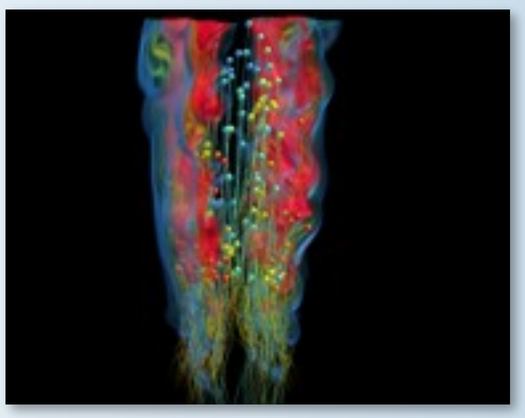
Center for Accelerated Application Readiness

WL-LSMS
Role of material disorder, statistics, and fluctuations in nanoscale materials and systems.



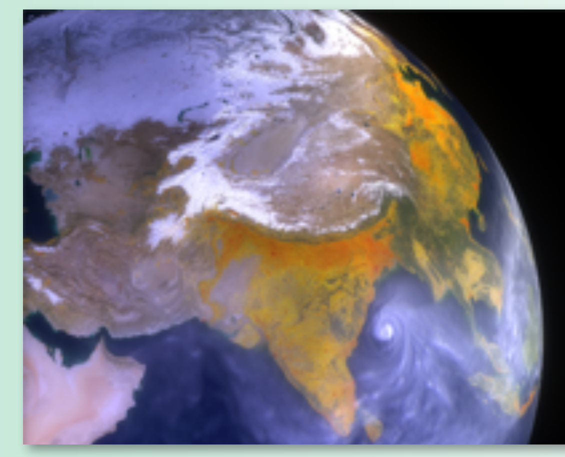


LAMMPS
Biofuels: An atomistic model of cellulose (blue) surrounded by lignin molecules comprising a total of 3.3 million atoms.

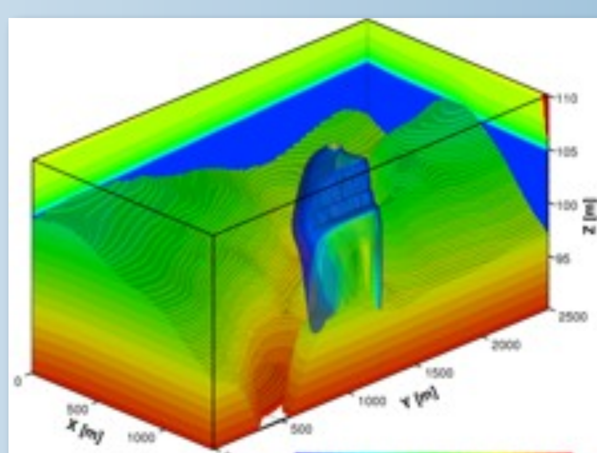


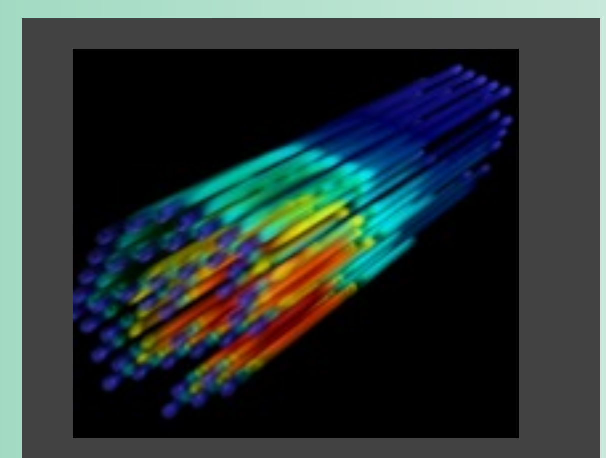
S3D
Understanding turbulent combustion through direct numerical simulation with complex chemistry.
.

CAM-SE
Answer questions about specific climate change adaptation and mitigation scenarios; realistically represent features like precipitation patterns/statistics and tropical storms.



PFLOTRAN
Stability and viability of large scale CO_2 sequestration; predictive containment groundwater transport.





Denovo
Discrete ordinates radiation transport calculations that can be used in a variety of nuclear energy and technology applications.

CAAR apps will form the vanguard of 'day-one' science on OLCF-3, but additional science teams will be granted friendly-user access as well (cf. our Petascale Early Science Period). Call for proposals will be forthcoming this summer.

CAAR Application Summary

Code	Description
CAM-HOMME	• Spectral finite element method • High leverage in physics packages • Scalable dynamical core of choice for future CCSM • Hard rating: Low compute intensity and high data movement in physics kernels
S3D	• DNS of combustion processes for specific fuels • Compressible Navier-Stokes flow solver for the full mass, momentum, energy and species conservation equations with structured grid written in F90 • Moderate rating: Complex rate equations, thermodynamics, and transport properties modules; no compute libraries used
LAMMPS	• Critical to development of alternative energy sources, including second-generation cellulosic ethanol • Easily broken up into components available to other MD codes • Broad open community MD code owned by a DOE national laboratory: Large user and developer groups • Moderate rating: Data non-locality due to calculation of long-range Coulomb force (common to all MD codes) – these changes will be made available as library

CAAR Application Summary (continued)

Code	Description
gWL-LSMS	• Enables first-principles studies of magnetic materials with broad relevance to DOE energy mission • Uses a workhorse approach (F77/90, C++, MPI) common to many applications • Straightforward rating: Main kernel based on dense linear algebra of complex numbers (LAPACK, CULA, MAGMA)
Denovo	• Key application for neutron transport and power distribution prediction in nuclear reactor cores • Moderate rating: Huge potential for exploiting untapped concurrency along "energy dimension" helps port, while heavy use of C++ and advanced programming models will tax GPU software and tool environment
PFLOTRAN	• Full featured finite element application with both structured and unstructured versions written in F90 • PETSc solver technology used extensively • Hard rating: Non-data locality caused by implicit nonlinear PDE solutions with indirect addressing and data movement caused by AMR (via SAMRAI)

CAAR Algorithmic Coverage

Code	FFT	Dense linear algebra	Sparse linear algebra	Particles	Monte Carlo	Structured grids	Unstructured grids
S3D		X	X	X		X	
CAM	X	X	X	X		X	
LSMS		X					
LAMMPS	X			X			
Denovo		X	X	X	X	X	
PFLOTRAN			X				X (AMR)

- Selected applications represented bulk of use for 6 INCITE allocations totaling 212M cpu-hours (2009)
  - –Represented 35% of 2009 INCITE allocations
  - –23% of 2010 INCITE allocations (in cpu-hours)

App	Science Area	Algorithm(s)	Grid type	Programming Language(s)	Compiler(s) supported	Communication Libraries	Math Libraries
CAM-HOMME	climate	spectral finite elements, dense & sparse linear algebra, particles	structured	F90	PGI, Lahey, IBM	MPI	Trilinos
LAMMPS	biology/materials	molecular dynamics, FFT, particles	N/A	C++	GNU, PGI, IBM, Intel	MPI	FFTW
S3D	combustion	Navier-Stokes, finite diff, dense & sparse linear algebra, particles	structured	F77, F90	PGI	MPI	None
Denovo	nuclear energy	wavefront sweep, GMRES	structured	C++, Fortran, Python	GNU, PGI, Cray, Intel	MPI	Trilinos, LAPACK, SuperLU, Metis
WL-LSMS	nanoscience	density functional theory, Monte Carlo	N/A	F77, F90, C, C++	PGI, GNU	MPI	LAPACK (ZGEMM, ZGTRF, ZGTRS)
PFLOTRAN	geoscience	Richards' equation coupled to transport and chemistry, finite-volume hydrodynamics	AMR	F90	PGI, GNU	MPI, SAMRAI	BLAS, PETSc

- **Algorithm and implementation coverage extends applicability well beyond the science domains immediately represented**

- **Much of the development work will also be pushed out to broader communities (e.g., in use of ChemKin)**

Tactics

- **Comprehensive team assigned to each app**
  - OLCF application lead
  - Cray engineer
  - NVIDIA developer
  - Other: other application developers, local tool/library developers
- **Particular plan-of-attack different for each app**
  - WL-LSMS – dependent on accelerated ZGEMM
  - CAM-HOMME – pervasive and widespread custom acceleration required
- **Multiple acceleration methods explored**
  - WL-LSMS – CULA, MAGMA, custom ZGEMM
  - CAM-HOMME – CUDA, PGI directives
  - Two-fold aim
    - **Maximum acceleration for model problem**
    - **<u>Determination of optimal, reproducible acceleration path for other applications</u>**
- **Constant monitoring of progress**
  - Status of each app discussed weekly

Application Teams

Application	OLCF Lead	Cray	NVIDIA	Science & Tools
S3D	Ramanan Sankaran	John Levesque	Gregory Ruetsch	Ray Grout (NREL)
WL-LSMS	Markus Eisenbach	Jeff Larkin Adrian Tate	Massimiliano Fatica Peng Wang	Yang Wang (PSC) Aurelian Rusanu (ORNL/UTK)
CAM-HOMME	Ilene Carpenter (NREL)	Jeff Larkin	Paulius Micikevicius	Matt Norman, Kate Evans, Rick Archibald, Jim Hack, Oscar Hernandez (ORNL) Mark Taylor (SNL) JF Lamarque, John Dennis (NCAR) Jim Rosinski (NOAA)
LAMMPS	Arnold Tharrington	Sarah Anderson	Peng Wang Scott Le Grande	Steve Plimpton, Paul Crozier (SNL) Mike Brown (ORNL) Axel Kohlmeyer (Temple) Mike Brown (OLCF)
Denovo	Wayne Joubert	Kevin Thomas	Cyril Zeller John Roberts	Tom Evans, Chris Baker (ORNL)
PFLOTRAN	Bobby Philip	Nathan Wichmann	Peng Wang	Peter Lichtner (LANL) Rebecca Hartmann-Baker (ORNL)

Complications

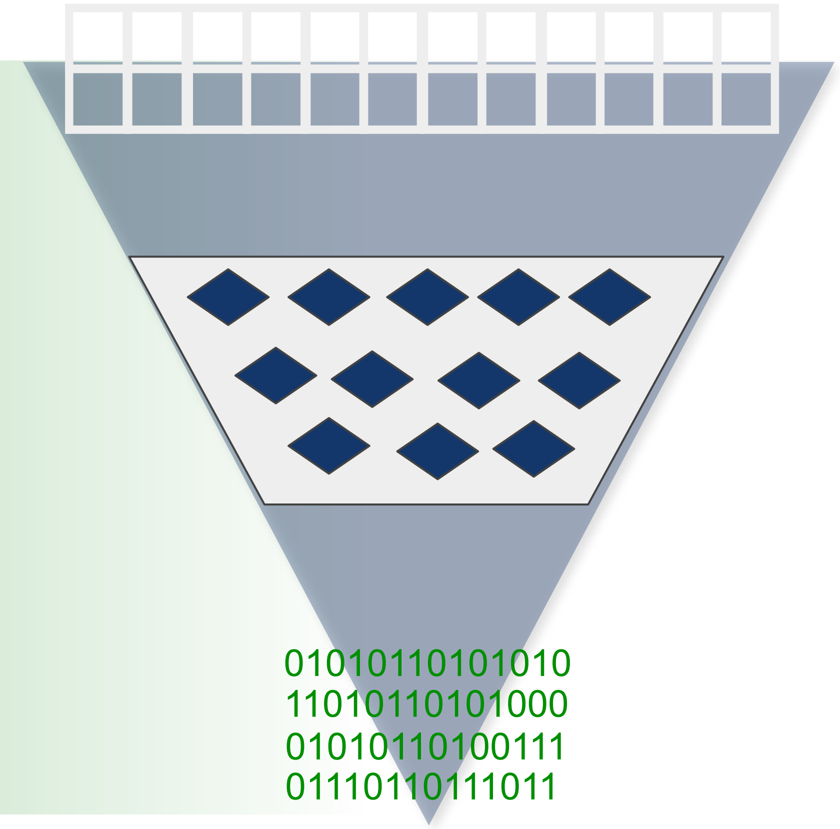
- **All of the chosen apps are under constant development**
  - Groups have, in many cases, already begun to explore GPU acceleration "on their own."
- **Production-level tools, compilers, libraries, etc. are just beginning to become available**
  - Multiple paths are available, with multifarious trade-offs
    - ease-of-use
    - (potential) portability
    - performance

What Are We Trying First?

- WL-LSMS
  - Primarily Library-based
- S3D
  - Directives and CUDA
- LAMMPS
  - CUDA
- CAM-SE
  - CUDA Fortran & Directives
- Denovo
  - CUDA
- PFLOTRAN
  - Directives

Hierarchical Parallelism

- **MPI parallelism between nodes (or PGAS)**
- **On-node, SMP-like parallelism via threads (or subcommunicators, or…)**
- **Vector parallelism**
  - SSE/AVX on CPUs
  - GPU threaded parallelism

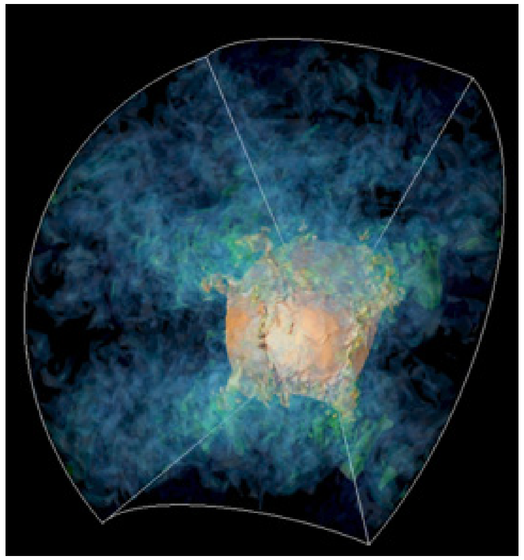


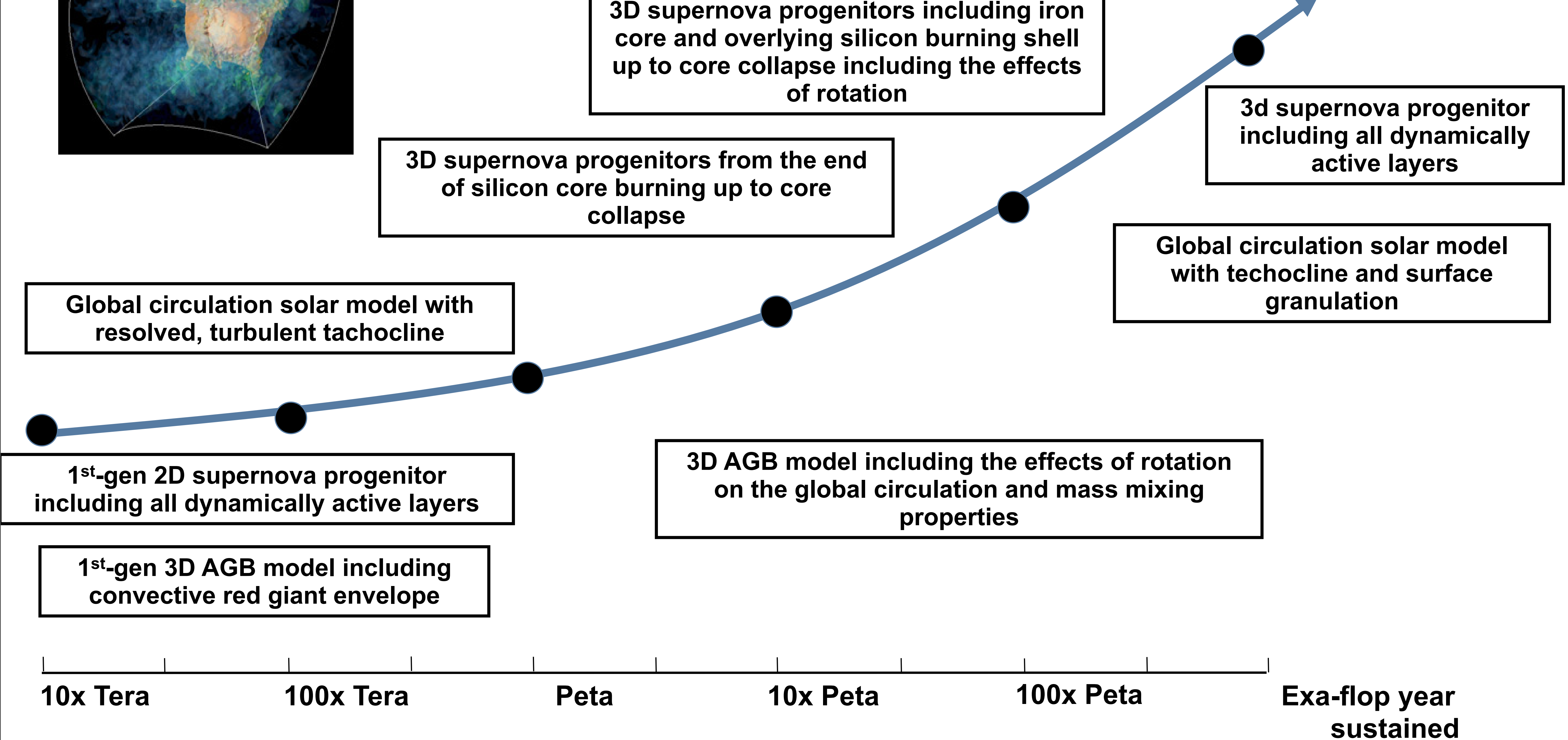
- **Exposure of unrealized parallelism is essential to exploit all near-future architectures.**
- **Uncovering unrealized parallelism and improving data locality improves the performance of even CPU-only code.**

Some Lessons Learned

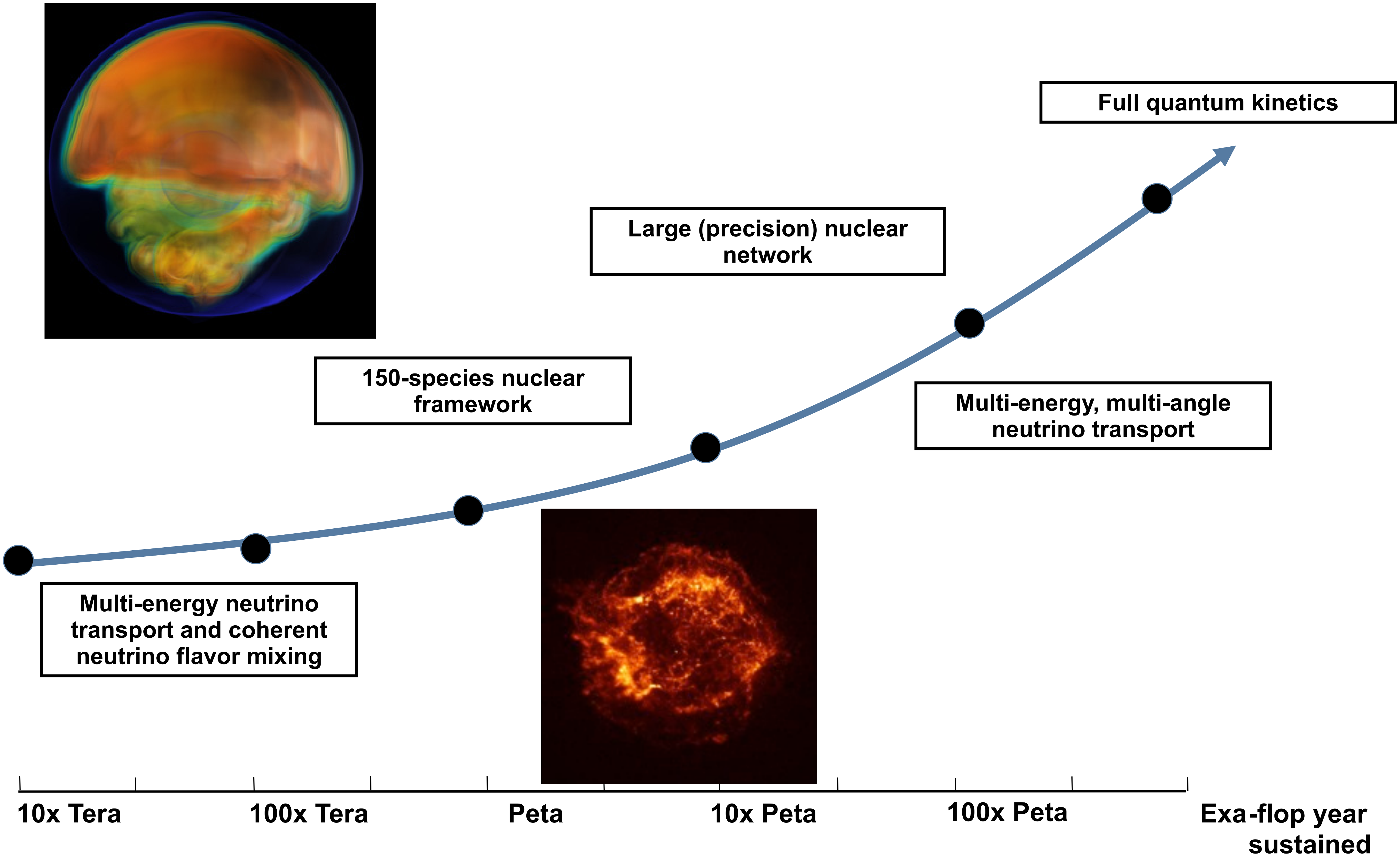
- **Exposure of unrealized parallelism is essential.**
  - Figuring out where is often straightforward
  - Making changes to exploit it is hard work (made easier by better tools)
  - Developers can quickly learn, e.g., CUDA and put it to effective use
  - A directives-based approach offers a straightforward path to portable performance
- **For those codes that already make effective use of scientific libraries, the possibility of continued use is important.**
  - HW-aware choices
  - Help (or, at least, no hindrance) to overlapping computation with device communication
- **Ensuring that changes are communicated back and remain in the production "trunk" is every bit as important as we initially thought.**
  - Other development work taking place on all CAAR codes could quickly make acceleration changes obsolete/broken otherwise
- **How much effort is this demanding?**
  - All 6 CAAR teams have converged (independently) to 2 ± 0.5 FTE-years

Stellar Evolution: The Sun and Other Stars

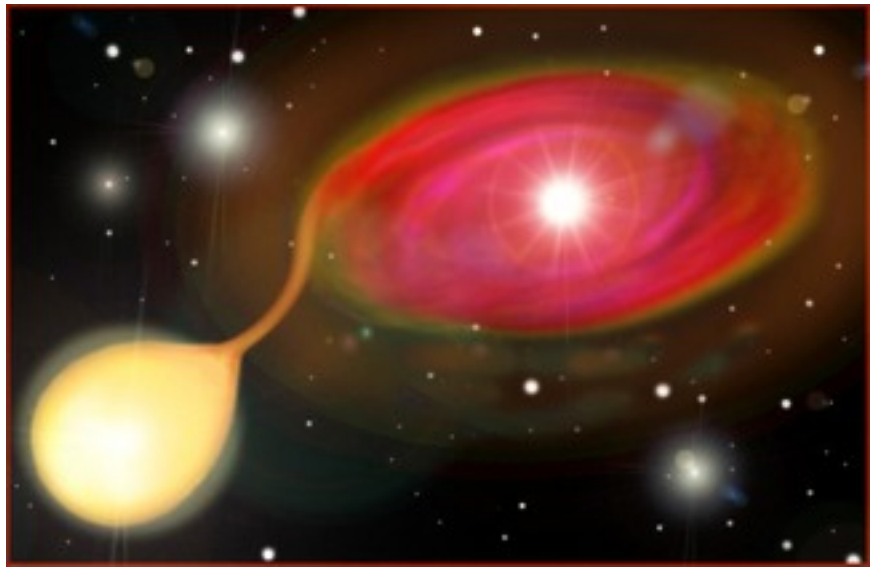


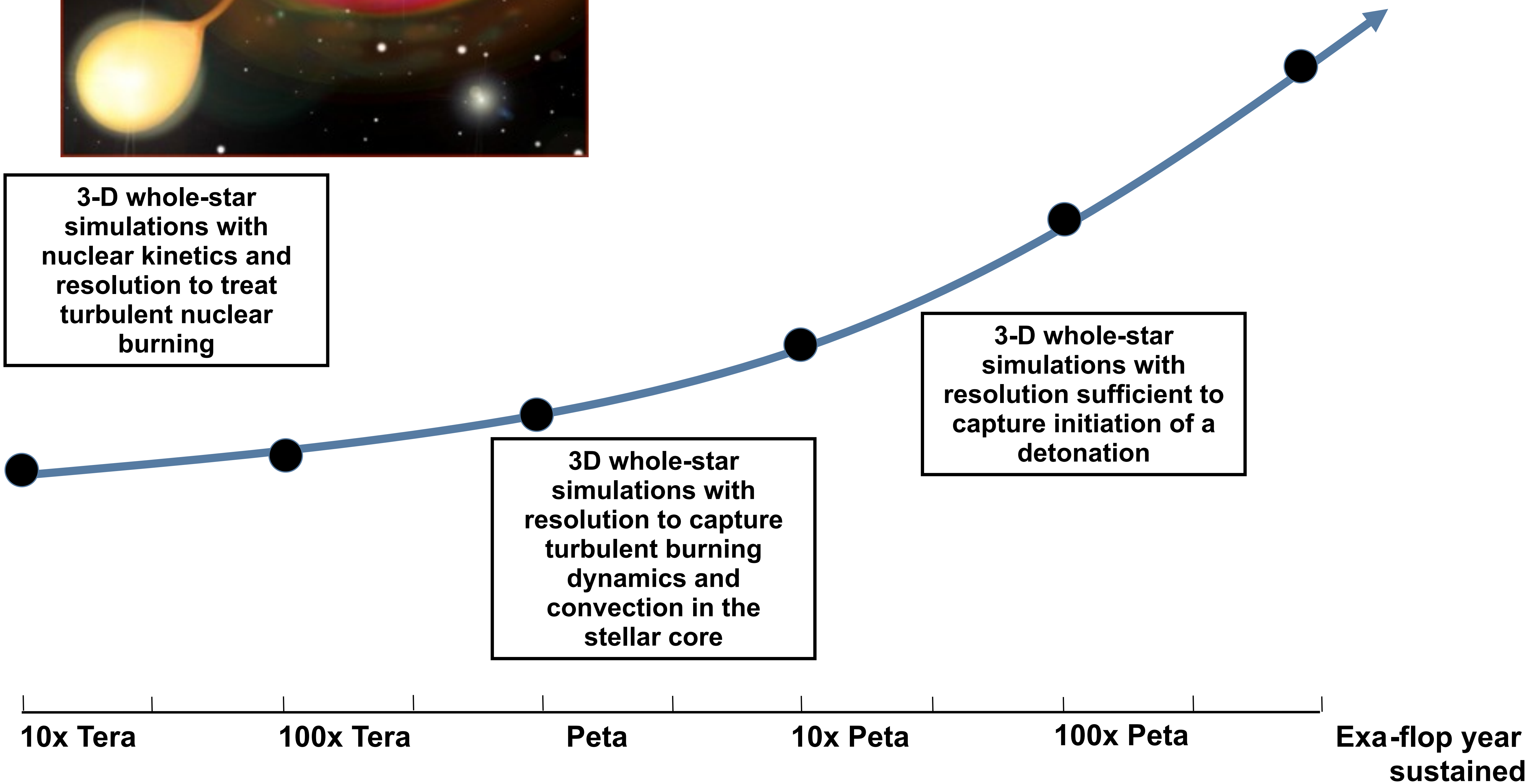


Core-Collapse Supernovae



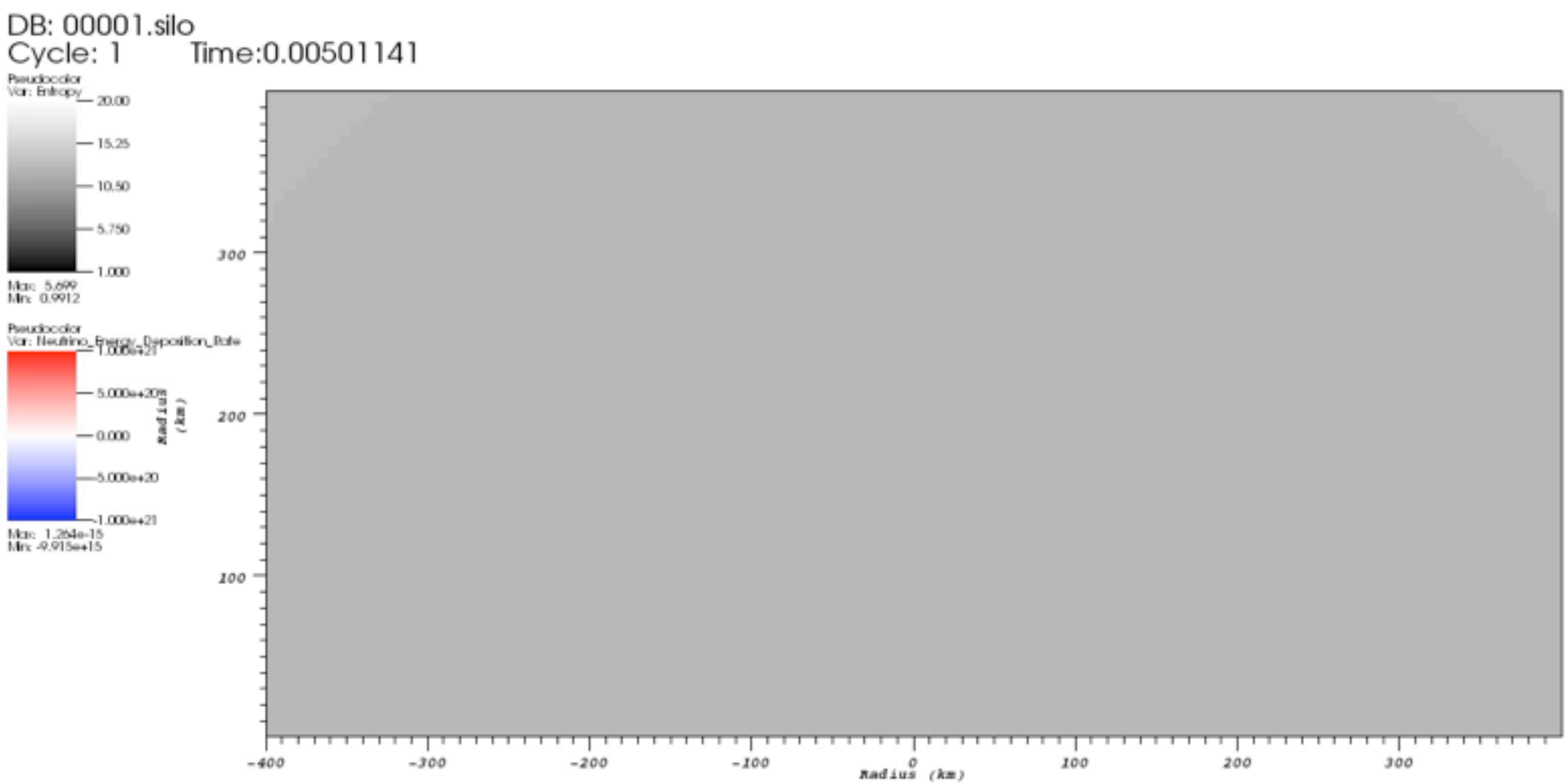
Thermonuclear Supernovae





Stellar Astrophysics provides a target-rich environment for these architectures

- Large number of DOF at each grid point
- Lots of opportunities to hide latency via multiphysics



Strong scaling with improved local physical fidelity is good, but not the whole answer.

- **Many problems (e.g. Type Ia SNe) are woefully underresolved**
- **Diminishing bytes/FLOP will limit spatial resolution (distributed memory)**
- **AMR will become even more essential**
  - Data locality becomes a problem
- **Task-based AMR systems**
  - cf. Uintah, MADNESS

Summary

- **We are not in the advent of exascale-like architectures, we are *in medias res*.**
- **Tools, compilers, etc. are becoming available to help make the transition.**
- **The specific details of the platforms matter much less than the overarching theme of hierarchical parallelism.**
- **Multiphysics simulations have unrealized parallelism to tap.**
  – Applications relying on, e.g., solution to large linear systems could also benefit from a task based approach.