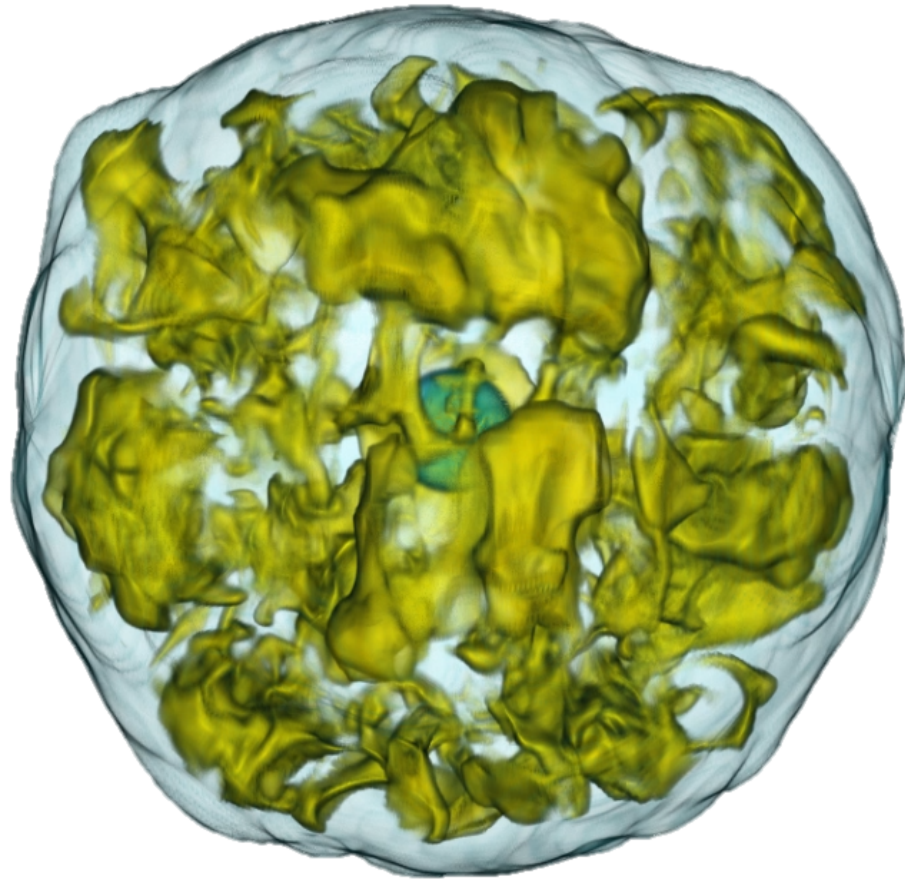


Mapping Matter Physics Baryon transport in the Early Universe



Bronson Messer

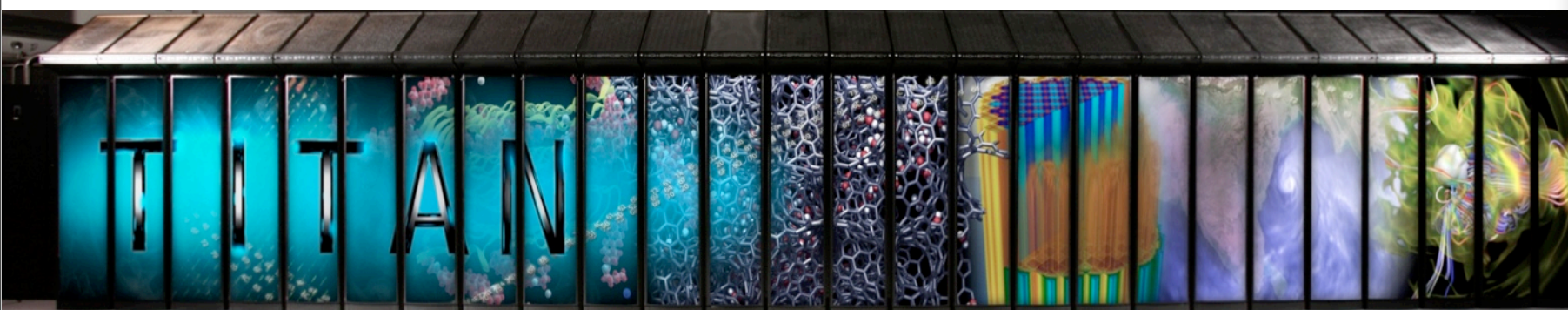
Scientific Computing
& Theoretical
Astrophysics Groups

Oak Ridge National
Laboratory

&

Department of
Physics & Astronomy

University of
Tennessee



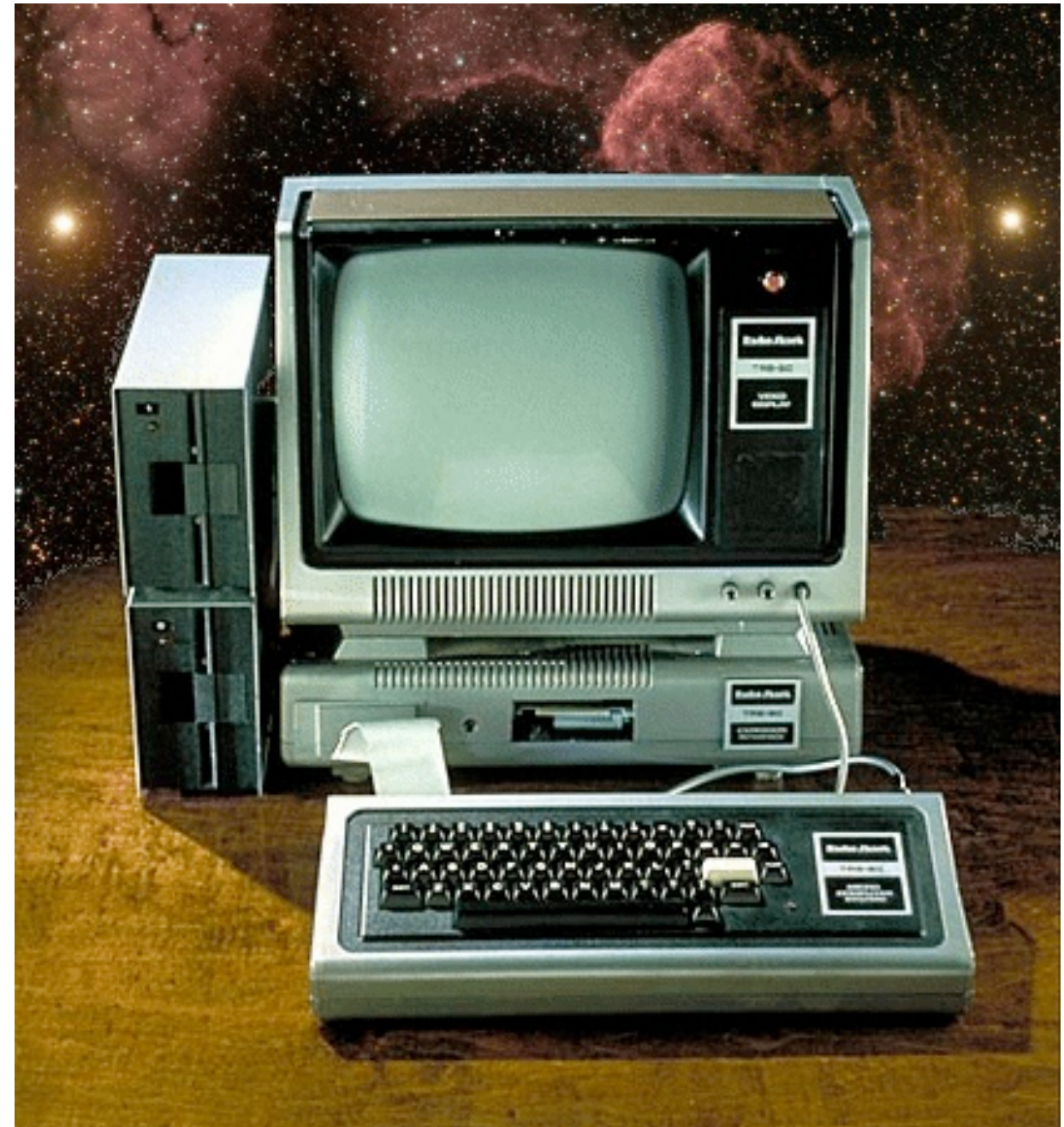
ORNL is managed by UT-Battelle
for the US Department of Energy



OAK RIDGE
National Laboratory

Strange things are afoot...

- We've done some parallel programming so far this week.
- Mostly confined to domain decomposition, implemented with MPI.
- This is 20th-century computing.



Nuclear astrophysics INCITE allocations from 2010 - present

- Average number of cpu-hours/yr ~152 M
- In aggregate, just less than 10% of the total available each year from 2010 - 2014
 - Allocations at NERSC are also above-average in size
 - PRAC is dominated by astrophysics, including considerable allocations for stellar astrophysics
- This excellent record is due to
 - the formulation of large, important problems
 - demonstrated ability to efficiently exploit the largest computational platforms
- Will this trend continue to the exascale?
Can we continue to solve big problems efficiently?



The Effects of Moore's Law and Slacking¹ on Large Computations

Chris Gottbrath, Jeremy Bailin, Casey Meakin, Todd Thompson,

J.J. Charfman

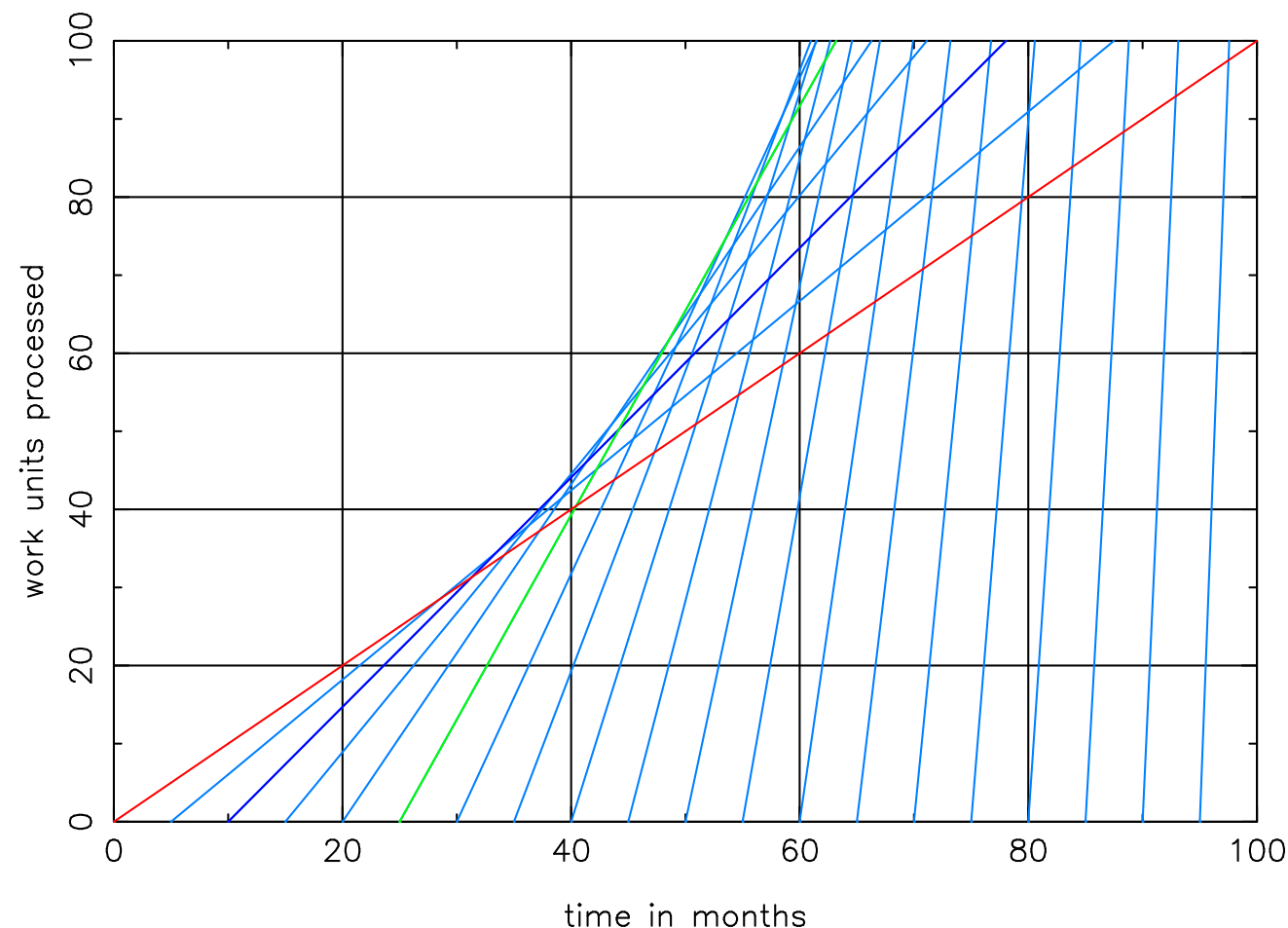
Steward Observatory, University of Arizona

¹This paper took 2 days to write

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.

work and slack in the context of moores law



astro-ph/9912202

OAK RIDGE | OAK RIDGE
National Laboratory | LEADERSHIP
COMPUTING FACILITY

DETECTION OF INTERSTELLAR BS IN THE CIRRUS DARK CLOUD OF THE NUMBBUM ASSOCIATION

I. AN INTUITIVE MODEL AND ITS SUBSEQUENT OBSERVATION

J.J. Charfman

deutsch english

suchen

On the Utter Irrelevance of LPL Graduate Students: An Unbiased Survey by Steward Observatory Graduate Students (2002)

[Charfman, J. J.](#), [Bsc, J. B.](#), [Eriksen, K. A.](#), [Knierman, K.](#), [Leistra, A.](#), [Mamajek, E.](#), [Monkiewicz, J.](#), [Moustakas, J.](#), [Murphy, J.](#), [Rigby, J.](#), [Young, P. A.](#)

Abstract

We present a new analysis of the irrelevance of Lunar and Planetary Laboratory (LPL) graduate students at the University of Arizona. Based on extensive Monte Carlo simulations we find that the actual number of useful results from LPL graduate students is ± 0.01 (5σ). Their irrelevance quotient far surpasses that of string theorists.

Details der Publikation

Download <http://arxiv.org/abs/astro-ph/0204041>
Archiv [arXiv \(United States\)](#)
Keywords Astrophysics
Typ text

1981) with broad-band alphabetical filters at the front of this book.

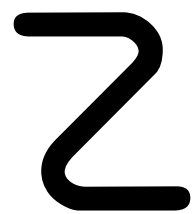
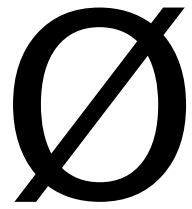
Alternate theories of BS formation on solid surfaces by Mayonnaise Hamburger (1977) have been considered, but only gas-produced BS can account for the amount reported here. However the yellow stuff reported by Hamburger (this volume) may well be mustard.

BS emission was generally observed in mornings and evenings, but was finally detected on Friday afternoon at a 1.8 σ level. Its lifetime was estimated to be approximately 5 days, but the level of emission was generally variable. The detection of this molecule has considerable impact on our understanding of Giant Gas Clouds which will be discussed in a later paper, but we want to note the most important of these at

645

*B. H. Andrew (ed.), Interstellar Molecules, 645–648.
Copyright © 1980 by the IAU.*

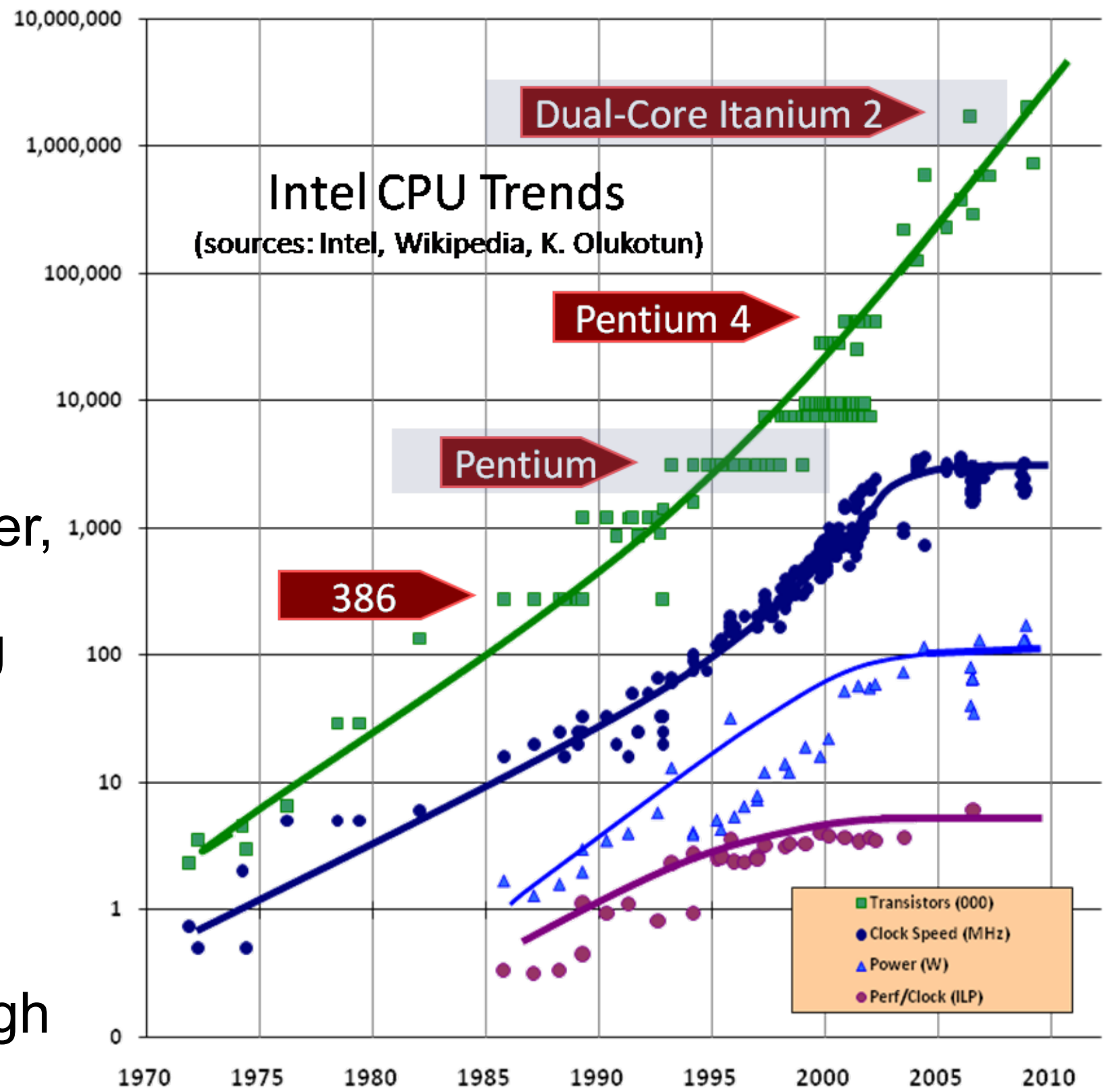
J. J. Charfman - The Snarky Nicolas Bourbaki of Arizona's Steward Observatory



dangerous bends, indeed,
are ahead...

Architectural Trends – No more free lunch

- CPU clock rates quit increasing in 2003
- $P = CV^2f$
Power consumed is proportional to the frequency and to the square of the voltage
- Voltage can't go any lower, so frequency can't go higher without increasing power
- Power is capped by heat dissipation and \$\$\$
- Performance increases have been coming through increased parallelism



Herb Sutter: Dr. Dobb's Journal:

<http://www.gotw.ca/publications/concurrency-ddj.htm>

Why are machines slowing down?

7 MegaWatts

- 1 MW-hour = 3.6×10^9 joules
- 1 gram TNT = 4184 joules (NIST)
- 1 jaguar-hour ~ 7 MW-hours ~ 2.5×10^{10} joules ~ 6 tons TNT
- 1 supernova simulation ~ 0.6 jaguar-months ~ 2.6 kton
- Trinity ~ 19 kton

But, you should care even if you don't use this much electricity...



What Burning Hot Laptops Can Do to Your Legs

5:10 AM - October 6, 2010 - By Jane McEntegart -
Source : Tom's Guide US

[f Like](#) 57 [f Send](#) [Twitter](#) 17 [+1](#) 0 [StumbleUpon](#) 1 [Share](#) 18

Hot laptops can lead to permanent skin discoloration from 'Toasted Skin Syndrome.'

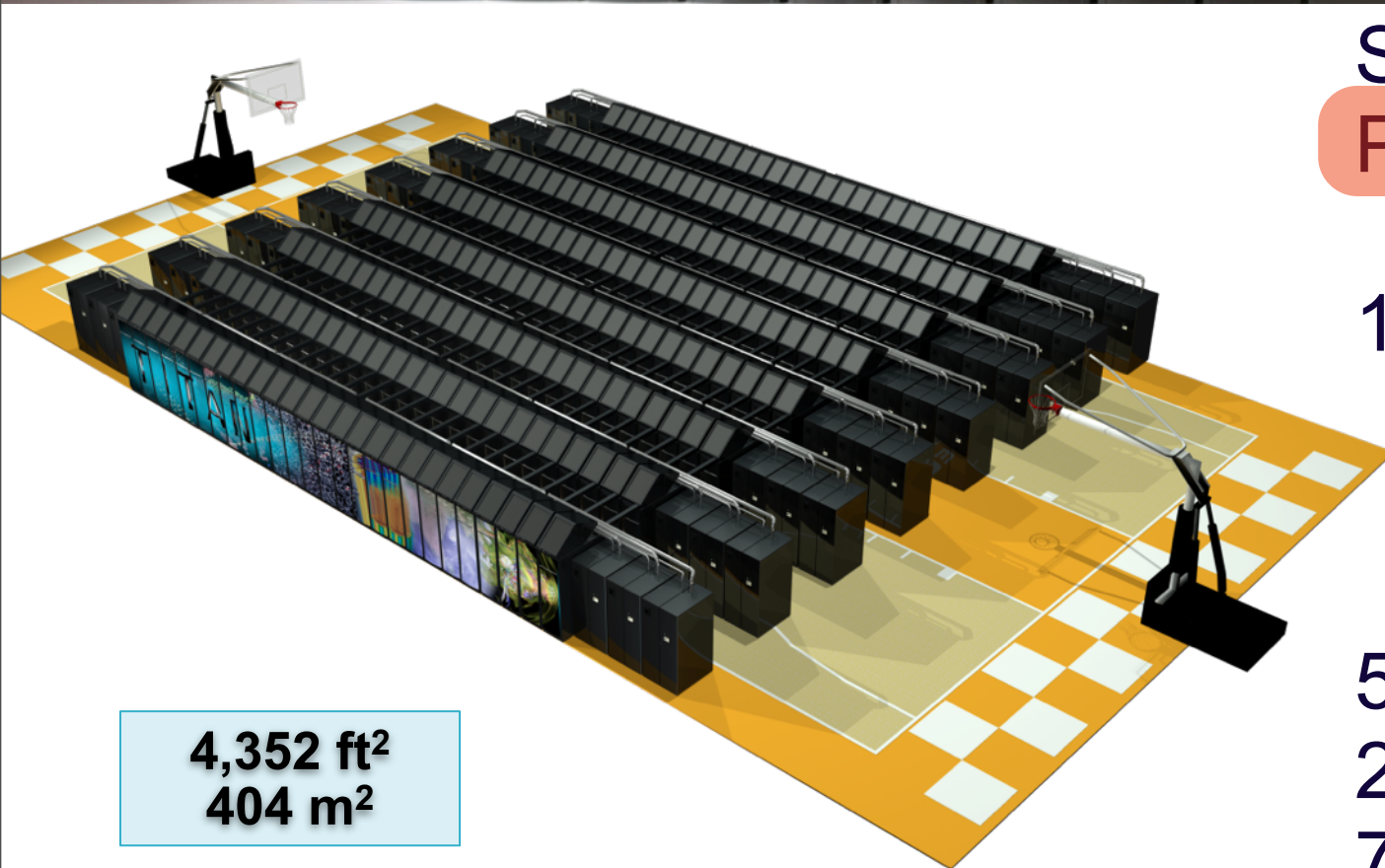
Most laptop owners are familiar with what happens when they use their computers on their laps. Things can get pretty warm and it doesn't take long for it to get uncomfortable. While most of us would move to the coffee table or stick a cushion under the laptop when things start to get sweaty, it seems not everyone takes the same measures and it's resulting in laptop-induced injuries.

The journal *Pediatrics* this week warned of the dangers of resting a hot laptop on your legs for extended periods of time and highlighted a couple of cases where it had resulted in erythema abigne, a skin condition the journal describes as "a reticular, pigmented, sometimes telangiectatic dermatosis that is caused by prolonged exposure to a heat or infrared source." Laptop-induced erythema abigne is now being referred to as "Toasted Skin Syndrome."



[Zoom](#)

ORNL's "Titan" Hybrid System: Cray XK7 with AMD Opteron and NVIDIA Tesla processors



4,352 ft²
404 m²

SYSTEM SPECIFICATIONS:

Peak performance of 27.1 PF >10x

24.5 GPU + 2.6 CPU

18,688 Compute Nodes each with:

16-Core AMD Opteron CPU

NVIDIA Tesla "K20x" GPU

32 + 6 GB memory

512 Service and I/O nodes

200 Cabinets

710 TB total system memory

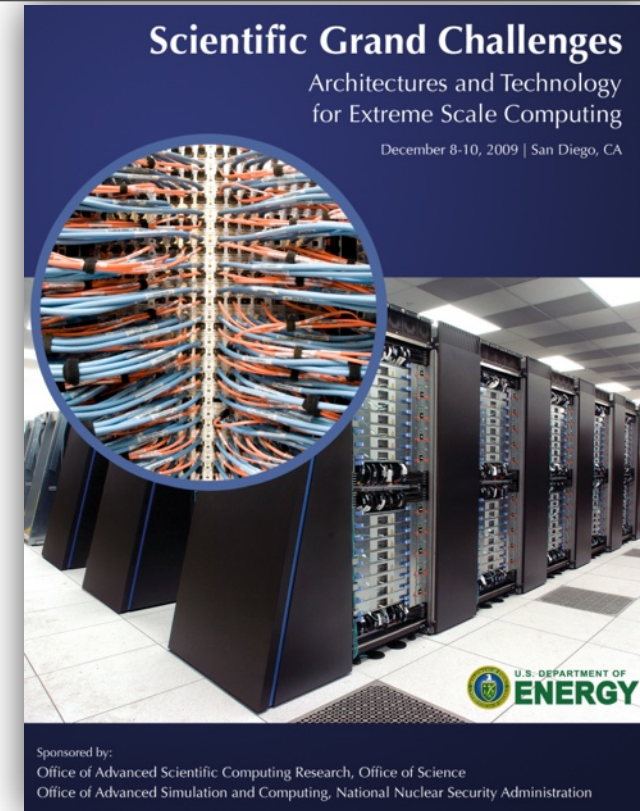
Cray Gemini 3D Torus Interconnect

8.8 MW peak power ~2x

What will the exascale look like?

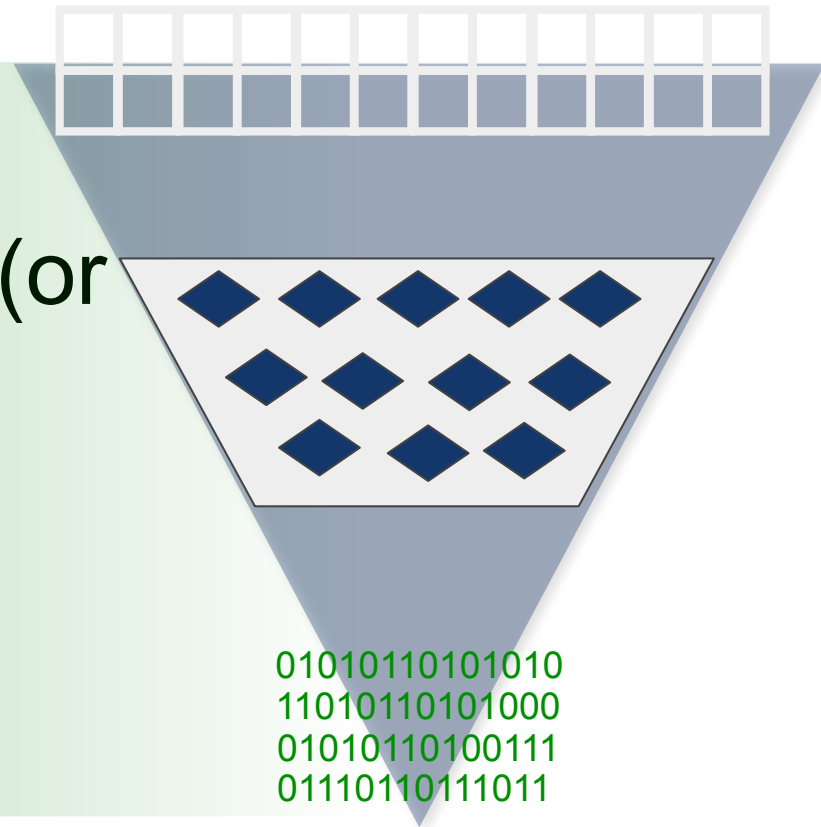
- “Node architectures are expected to change dramatically in the next decade, becoming more hierarchical and heterogeneous.”
- “. . . computer companies are dramatically increasing on-chip parallelism to improve performance. The traditional doubling of clock speeds every 18 to 24 months is being replaced by a doubling of cores or other parallelism mechanisms.”
- “Systems will consist of one hundred thousand to one million nodes and perhaps as many as a billion cores.”

Architectures and Technology for Extreme Scale Computing, Workshop Report, 2009;
<http://www.er.doe.gov/ascr/ProgramDocuments/Docs/Arch-TechGrandChallengesReport.pdf>



Hierarchical Parallelism

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

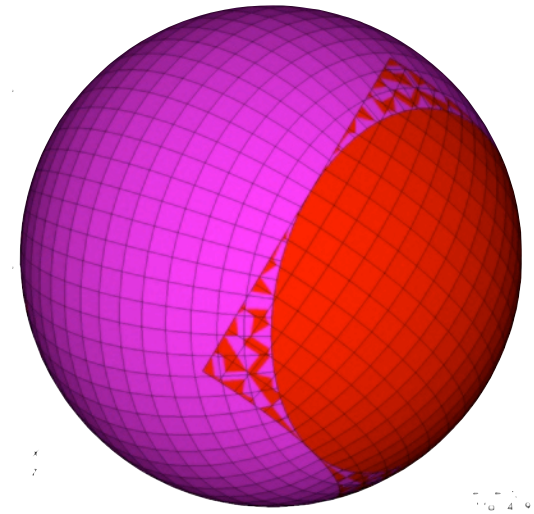


- **Exposure of unrealized parallelism is essential to exploit all near-future architectures.**
- **Uncovering unrealized parallelism and improving data locality improves the performance of even CPU-only code.**
- **Experience with vanguard codes at OLCF suggests 1-2 person-years is required to “port” extant codes to GPU platforms.**
 - **Likely less if begun today, due to better tools/compiler**

So how should we program for these new systems?

- What to do – Good Threading (OpenMP)
 - Must do high level threading
 - Thread must access close shared memory rather than distant shared memory
 - Load Balancing
- What to do – Good Vectorization
 - Vectorization advantage allows for introducing overhead to vectorize
 - Vectorization doesn't rely on having HW that is described as “vector”

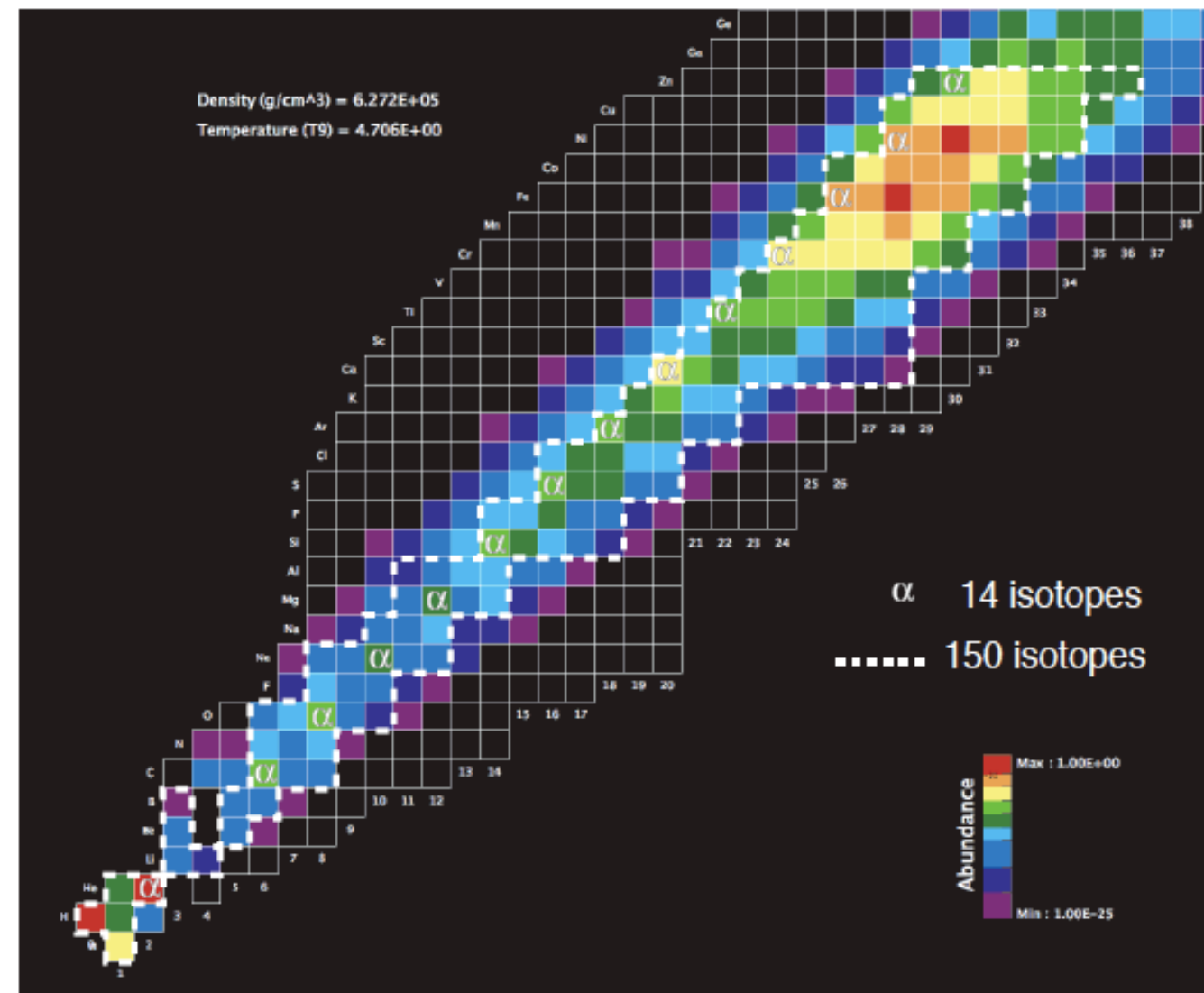
Good news! Stellar astrophysics tends to have a lot of unrealized parallelism at present



- **Simulation codes for stellar evolution and explosions**
 - Exemplars of “multiphysics application codes”
 - Typically many degrees-of-freedom per spatial grid point
 - radiation transport
 - nuclear burning
 - Spatial domains typically parallelized via domain decomposition

- **Related, computationally-intensive topics will, perhaps, have to work harder to identify additional parallelism outside of large stellar simulations, but plenty of opportunity exists.**

- high-density physics
- nuclear structure and reactions



- A common directive programming model for today's GPUs
 - Announced at SC11 conference
 - Offers portability between compilers
 - Drawn up by: NVIDIA, Cray, PGI, CAPS
 - Multiple compilers offer portability, debugging,
 - permanence
 - Works for Fortran, C, C++
 - Standard available at www.OpenACC-standard.org
 - Initially implementations targeted at NVIDIA GPUs
- Current version: 1.0 (November 2011)
- Compiler support:
 - Cray CCE: complete 1.0
 - PGI Accelerator: complete 1.0 + extensions
 - CAPS: complete 1.0 + extensions



Why use OpenACC Directives?

- Productivity
 - Higher level programming model
 - *a la* OpenMP
- Portability
 - ignore directives, portable to the host
 - portable across different accelerators
 - *performance portability*
- Performance feedback

OpenCL

CUDA C

```

29 workSize[i] * WA, 0, NULL, NULL);
30
31 // create OpenCL buffer on device that will be initialized from the host memory on first use
32 // on device
33 d_B[i] = clCreateBuffer(cxGPUContext, CL_MEM_READ_ONLY | CL_MEM_COPY_HOST_PTR,
34 mem_size_B, h_B_data, NULL);
35
36 // Output buffer
37 d_C[i] = clCreateBuffer(cxGPUContext, CL_MEM_WRITE_ONLY, workSize[i] * WC * sizeof(float), NULL, NULL);
38
39 // set the args values
40 clSetKernelArg(multiplicationKernel[i], 0, sizeof(cl_mem), (void *) &d_C[i]);
41 clSetKernelArg(multiplicationKernel[i], 1, sizeof(cl_mem), (void *) &d_A[i]);
42 clSetKernelArg(multiplicationKernel[i], 2, sizeof(cl_mem), (void *) &d_B[i]);
43 clSetKernelArg(multiplicationKernel[i], 3, sizeof(float) * BLOCK_SIZE * BLOCK_SIZE, 0);
44 clSetKernelArg(multiplicationKernel[i], 4, sizeof(float) * BLOCK_SIZE * BLOCK_SIZE, 0);
45
46 int by = blockIdx.y;
47 int tx = threadIdx.x;
48 int ty = threadIdx.y;
49 if(i+1 < ciDeviceCount)
50 workOffsetInBlock = workOffsetInBlock + workSize[i];
51 int aBegin = aBegin + WA - 1;
52 int aStep = BLOCK_SIZE;
53 // Execute Multiplication on all GPUs in parallel
54 size_t localWorkSize[] = {BLOCK_SIZE, BLOCK_SIZE};
55 size_t globalWorkSize[] = {shrRoundUp(BLOCK_SIZE, WC), shrRoundUp(BLOCK_SIZE, workSize[0])};
56 // Launch kernels on devices
57 for(unsigned int i = 0; i < ciDeviceCount; i++)
58 {
59     shared float As[BLOCK_SIZE][BLOCK_SIZE];
60     shared float Bs[BLOCK_SIZE][BLOCK_SIZE];
61     // Multiplication with non-blocking execution
62     BS[ty + tx * WC] = B[b + WB * ty + tx];
63     globalWorkSize[1] = shrRoundUp(BLOCK_SIZE, workSize[i]);
64     clEnqueueNDRangeKernel(commandQueue[i], multiplicationKernel[i], 2, 0, globalWorkSize, localWorkSize,
65     Csub += AS(ty, k) * BS(k, tx), 0, NULL, &GPUExecution[i]);
66     _syncthreads();
67 }
68 for(unsigned int i = 0; i < ciDeviceCount; i++)
69 {
70     C[C + WB * ty + tx] = Csub;
71 }
72 clFinish(commandQueue[i]);
73 }
74 void
75 computeMM0_saxpy(float C[ciDeviceCount][WB], float A[ciDeviceCount][WA], float B[ciDeviceCount][WB],
76 int hA, int hB, int hC)
77 {
78     // Non-blocking copy of result from device to host
79     clEnqueueReadBuffer(commandQueue[i], d_C[i], CL_FALSE, 0, workSize[i] * WC * sizeof(float) * workSize[i],
80     unsigned int size_C = WC * hC, workOffset[i], 0, NULL, &GPUOne[i]);
81     float *d_A, *d_B, *d_C;
82     // GPU sync with CPU
83     cudaMalloc((void**) &d_A, mem_size_A);
84     cudaMalloc((void**) &d_B, mem_size_B);
85     cudaMalloc((void**) &d_C, mem_size_C);
86     cudaMemcpy(d_A, h_A, mem_size_A, cudaMemcpyHostToDevice);
87     cudaMemcpy(d_B, h_B, mem_size_B, cudaMemcpyHostToDevice);
88     for(unsigned int i = 0; i < ciDeviceCount; i++)
89     {
90         dim3 threads(BLOCK_SIZE, BLOCK_SIZE);
91         dim3 grid(WC / threads.x, HC / threads.y);
92         clReleaseMemObject(d_A[i]);
93         clReleaseMemObject(d_B[i]);
94         clReleaseMemObject(d_C[i]);
95         clReleaseEvent(GPUExecution[i]);
96         clReleaseEvent(GPUOne[i]);
97         cudaFree(d_A);
98         cudaFree(d_B);
99         cudaFree(d_C);
100     }
101 }
102
103 __kernel void
104 matrixMul( __global float* C, __global float* A, __global float* B,
105 __local float* As, __local float* Bs)
106 {
107     int bx = get_group_id(0), tx = get_local_id(0);
108     int by = get_group_id(1), ty = get_local_id(1);
109     int aEnd = WA * BLOCK_SIZE * by + WA - 1;
110
111     float Csub = 0.0f;
112
113     for (int a = WA*BLOCK_SIZE*by, b = BLOCK_SIZE * bx;
114 a <= aEnd; a += BLOCK_SIZE, b += BLOCK_SIZE*WB) {
115     As[tx + ty * BLOCK_SIZE] = A[a + WA * ty + tx];
116     Bs[tx + ty * BLOCK_SIZE] = B[b + WB * ty + tx];
117     barrier(CLK_LOCAL_MEM_FENCE);
118     for (int k = 0; k < BLOCK_SIZE; ++k)
119     Csub += As[k + ty * BLOCK_SIZE]*Bs[tx + k * BLOCK_SIZE] ;
120     barrier(CLK_LOCAL_MEM_FENCE);
121 }
122 C[get_global_id(1) * get_global_size(0) + get_global_id(0)] = Csub;
123 }
124 }
125 }

```

OpenAcc



Risk factors

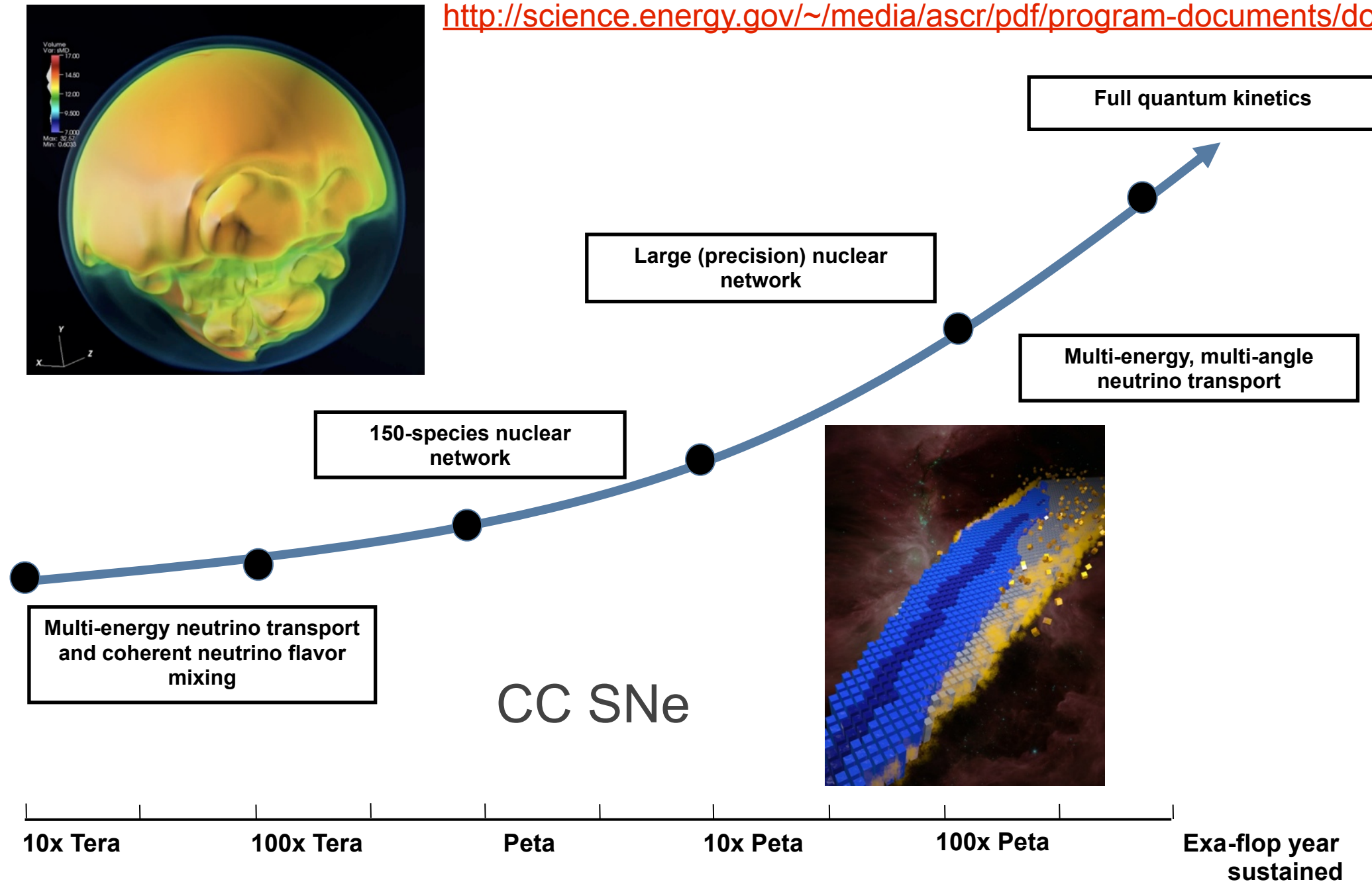
- Will there be machines to run my OpenACC code on?
 - Now? Lots of Nvidia GPU accelerated systems
 - Cray XK7s: CSCS tödi, HLRS hermit, ORNL titan...
 - Lots of other GPU machines in Top100 (OpenACC is multi-vendor)
 - Future? OpenACC can be targeted at other accelerators
 - PGI and CAPS already target Intel Xeon Phi, AMD GPUs
 - Plus you can always run on CPUs using same codebase
- Will OpenACC continue?
 - Support? Cray and PGI (at least) are committed to support OpenACC
 - Lots of big customer pressure to continue to run OpenACC
 - Develop? OpenACC v2.0 standard out
 - Lots of new partners joined committee at end of last year
- Will OpenACC be superseded by something else?
 - Auto-accelerating compilers? If only!
 - Never really managed it for threading real HPC applications on the CPU
 - Data locality adds to the challenge
 - **OpenMP accelerator directives? OpenACC work not wasted**
 - Very similar programming model; can transition easily
 - Cray (co-chair), PGI very active in OpenMP accelerator subcommittee

Posited Exascale Specs

| System attributes | 2010 | 2018 | | 2023? | |
|----------------------------|----------|-------------------|----------|-------------------|-----------|
| | | “2015” | | “2018” | |
| System peak | 2 PF | 200 PF/s | | 1 Exaflop/s | |
| Power | 6 MW | 15 MW | | 20 MW | |
| System memory | 0.3 PB | 5 PB | | 32–64 PB | |
| Node performance | 125 GF | 0.5 TF | 7 TF | 1 TF | 10 TF |
| Node memory BW | 25 GB/s | 0.1 TB/s | 1 TB/s | 0.4 TB/s | 4 TB/s |
| Node concurrency | 12 | O(100) | O(1,000) | O(1,000) | O(10,000) |
| System size (nodes) | 18,700 | 50,000 | 5,000 | 1,000,000 | 100,000 |
| Total node interconnect BW | 1.5 GB/s | 150 GB/s | 1 TB/s | 250 GB/s | 2 TB/s |
| MTTI | day | O(1 day) | | O(1 day) | |

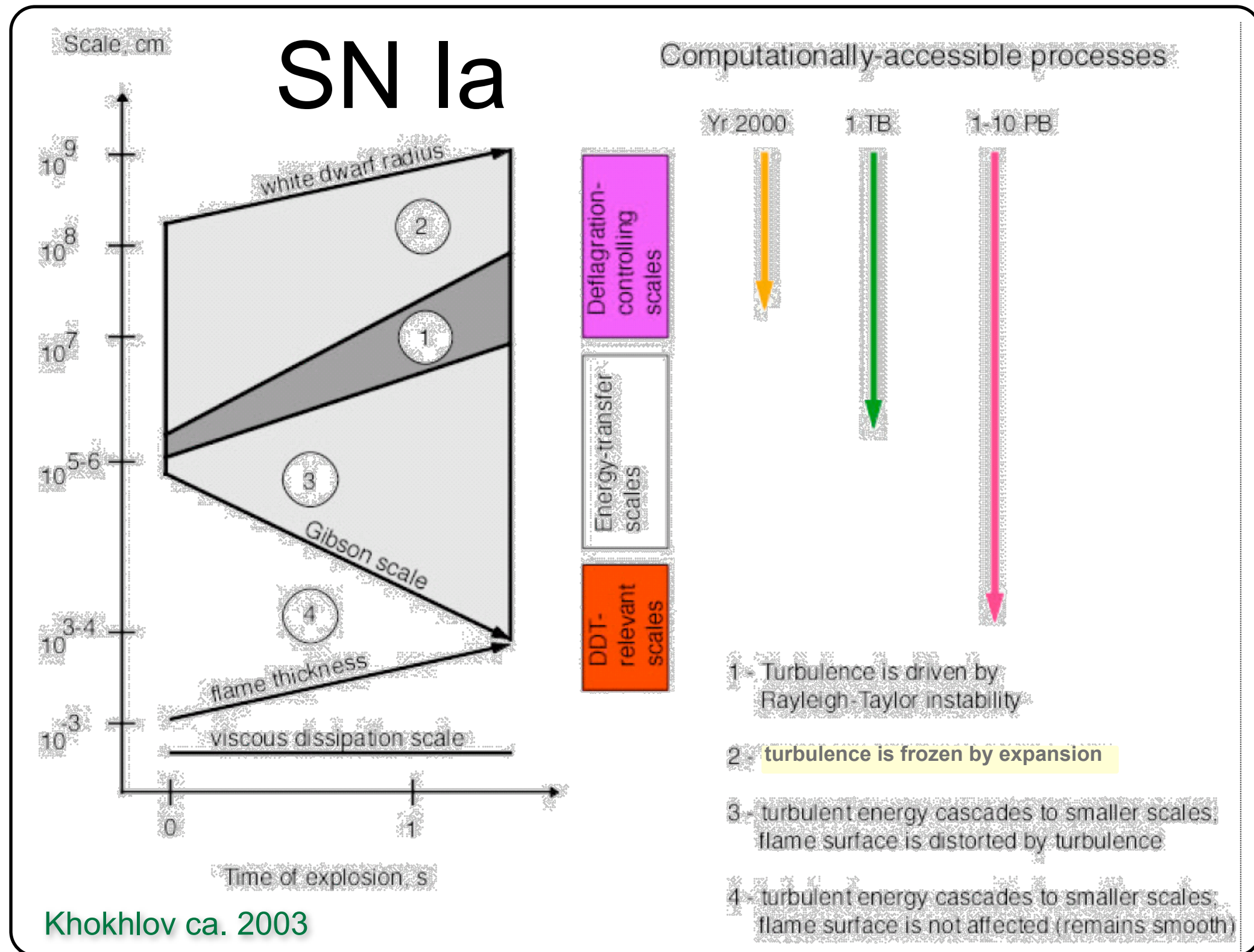
Example goals/highlights from NP Exascale Workshop report (2009)

http://science.energy.gov/~media/ascr/pdf/program-documents/docs/Np_report.pdf



All of these goals are attainable, but will require new algorithms and implementations to bridge the gap to the posited architectures.

Achieving high spatial (or phase-space, etc.) resolution will be very difficult.



Total memory on the **entire** exascale system will be O(10 PB)

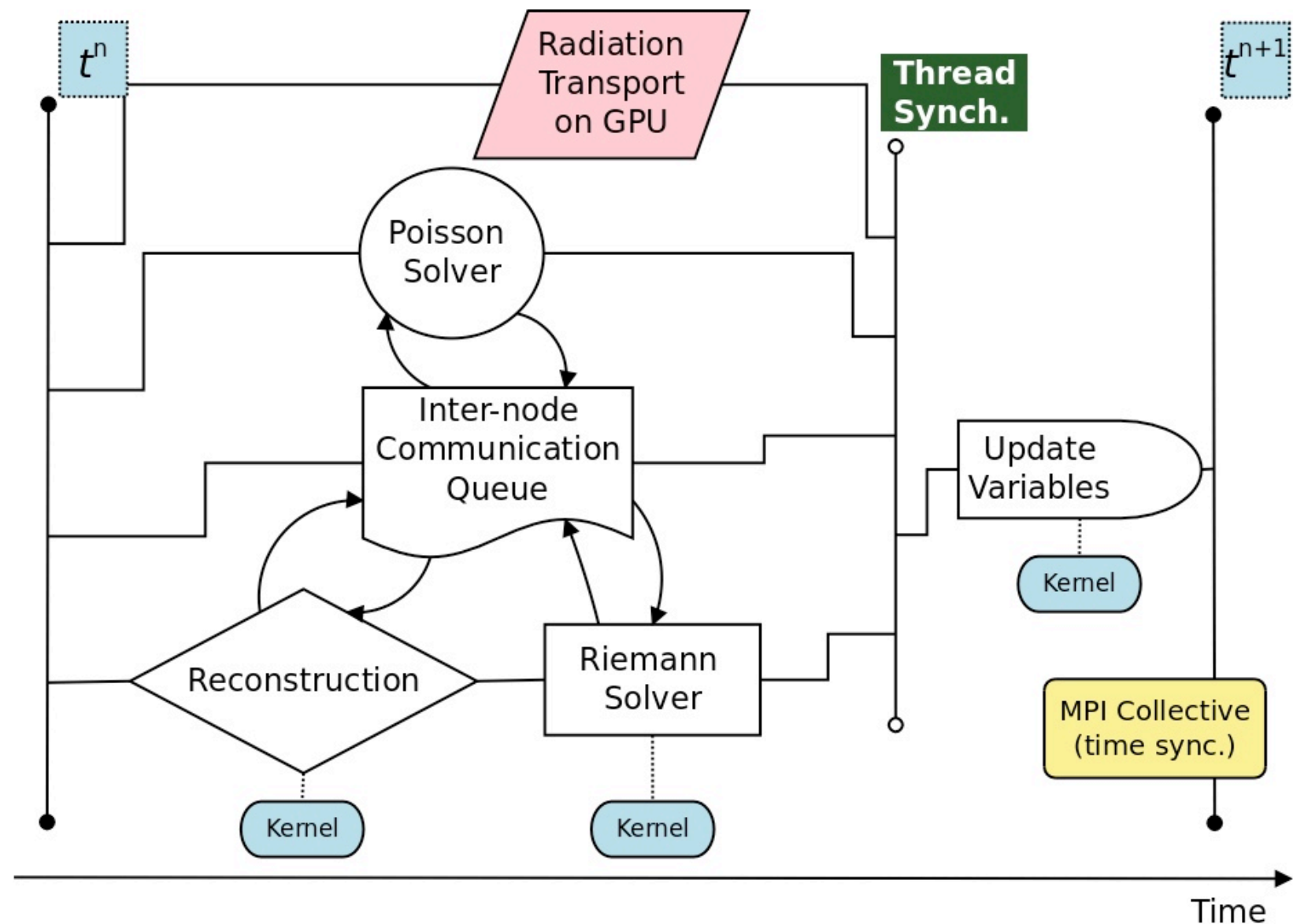
Locally bulk-synchronous programming model is not a viable path for maximum performance on these new platforms

- FLOP/s are cheap and moving data is expensive.
- Even perfect knowledge of resource capabilities at every moment and perfect load balancers will not rescue billion-thread SPMD implementations of PDE simulations.
- Cost of rebalancing frequently is too large, but the Amdahl penalty of failing to rebalance is fatal.
- To take full advantage of asynchronous algorithms, we need to develop greater expressiveness in scientific programming.
 - Create separate threads for logically separate tasks, whose priority is a function of algorithmic state, not unlike the way a time-sharing OS works.
 - Join priority threads in a directed acyclic graph (DAG), a task graph showing the flow of input dependencies; fill idleness with noncritical work or steal work.

Comments taken directly from keynote address by David Keyes at EU-US HPC Summer School, June 2012

Asynchronous execution models via task scheduling

- Examples exist already in other domains
 - MAGMA (linear algebra)
 - MADNESS (DFT)
 - Uintah (terrestrial combustion)
- Operator-split physics modules become “tasks” associated with execution threads



Will the exascale (or before) machine be primarily a “strong-scaling” platform?

- Memory constraints provide a hard ceiling for spatial resolution and number of unknowns.
 - bytes/FLOP goes down by an order of magnitude
- Simulations will be certainly be larger, but likely not as large as one would expect if scaling with FLOPs is assumed.
 - no more than $\sim 10x$ the number of MPI ranks?
 - this connotes no more than factors of ~ 2 in resolution in each dimension for 3D
- OK: considerable understanding can be realized by fully exploring parameter space.
 - progenitor mass, rotation, metallicity
 - transport approximations, additional physics

Simulation, code, and data management become even harder



- Revision control, regression testing, viz, workflow...

The screenshot displays the Bellerophon Visualization On-Demand interface. At the top, there are navigation buttons: Regression Test Explorer, Visualization Explorer, Visualization On-Demand (highlighted), SVN Statistics On-Demand, Important Links & Documentation, and Log Out and Quit Bellerophon.

The main content area is divided into several sections:

- Visualization Sets Table:** A table listing simulation sets with columns for Index, Chimera Sim ID, Created By, Creation Date, Progenitor Source, Progenito..., and Resolution. The table contains 14 rows, with the 12th row (Data3_SASI_v_diff_03_alt_ptrb) selected.
- Search Filter:** A search bar with a dropdown menu for Index and a text input field for Chimera Simulation ID.
- Visualization Set Notes:** A text area containing the note: "Accretion shock simulation with an EOS gamma of 4/3, and 'wiggle' parameter of 0.03. Convection is set random velocity perturbations with a min to max of -.05 ϵ_{sound} to 0.05 ϵ_{sound} ."
- Visualization Set Details:** A section showing details for the selected visualization set (Index 12):
 - Visualization Set Index = 12
 - Chimera Simulation ID = Data3_SASI_v_diff_03_alt_ptrb
 - Created By = Lingerfelt, Eric
 - Creation Date = 09/18/10
 - Progenitor Source = No Progenitor Source
 - Progenitor Mass = -1
 - Resolution = 252 x 252
 - Animation Index = 20
 - Size [pixels] = 1280 x 710
 - Colortable = orangehot
 - Zoom [km] = -300,300,0,300
 - Colortable Range = 1.8
- Animation Information:** A section with buttons for Save, Print, and Copy.
- Animation Viewer:** A large window showing a simulation of a star's surface. The title is "Data3_SASI_v_diff_03_alt_ptrb (12) → Entropy (s) (20)". The viewer displays a color map of the star's surface with a color scale from 1.00 to 4.00. The axes are labeled "radius (km)" and "Time: 0.1415". The viewer includes playback controls (play, stop, back, forward) and options for "Fit to Window" and "Full Screen Mode".

At the bottom of the interface, there are buttons for "Select Different Option", "Select New Viz Sets", "Download Frame", and "Download Movie".

Summary

- **The future is now! Computers are not getting faster from the perspective of a nuclear astrophysicist. They are only getting “wider.” This is true on the world’s largest supercomputers and on your laptop!!!!**
- **The Xeon(Phi)/GPU/BG/Q choice is no choice at all. They are all versions of a single narrative.**
- **Stellar astrophysics is rife with unrealized parallelism, but architectural details and memory (i.e. cost, power) constraints will present considerable challenges.**
- **Bulk-synchronous execution is a terrible way to try to exploit near-future architectures. A new programming model will require considerably more effort than a simple multi/many-core port.**
- **Managing large simulations is something we can barely do now, but how about managing 1000’s, 10’s of thousands, or 100’s of thousands of simulations? We should not expect to rely on solutions to be thrown over the fence from developers in other communities.**

