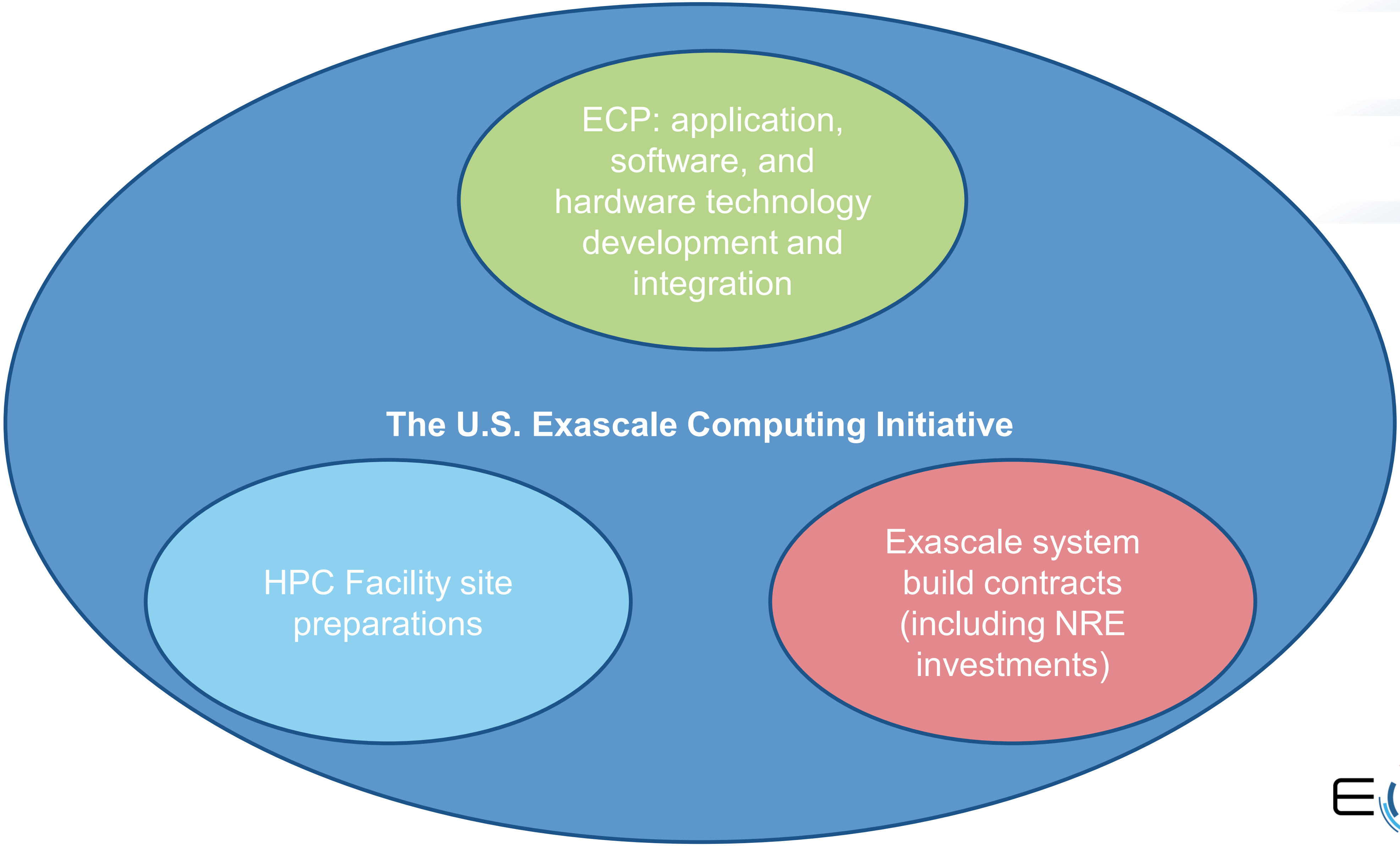
Preparing Applications for Next-Generation HPC Architectures

Andrew Siegel

Argonne National Laboratory

EXASCALE COMPUTING PROJECT

Exascale Computing Project (ECP) is part of a larger US DOE strategy

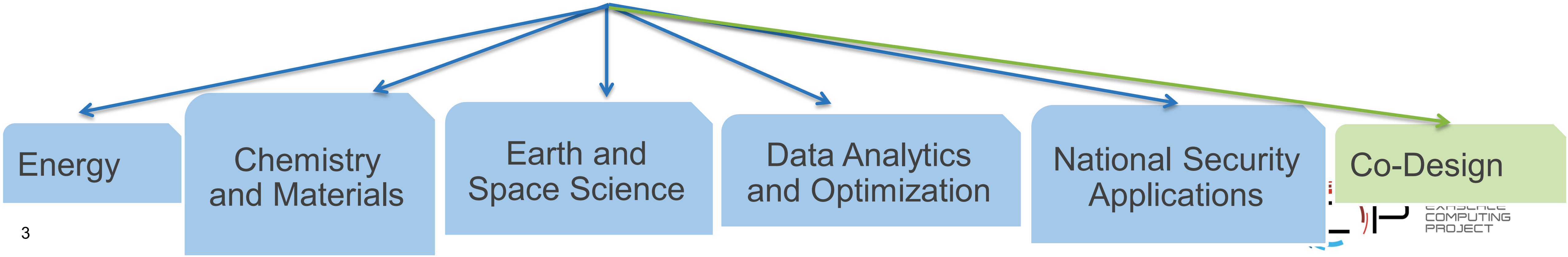


Exascale Computing Project

- Department of Energy project to develop usable exascale ecosystem

- Exascale Computing Initiative (ECI)
  1. 2 Exascale platforms (2021)
  2. Hardware R&D
  3. System software/middleware
  4. 25 Mission critical application projects

Exascale Computing Project (ECP)



Pre-Exascale Systems

Exascale Systems

2013	2016	2018	2020	2021-2023

Argonne
IBM BG/Q
Open

ORNL
Cray/NVidia K20
Open

LLNL
IBM BG/Q
Secure

Argonne
Intel/Cray KNL
Open

LBNL
Cray/Intel Xeon/KNL
Open

LANL/SNL
Cray/Intel Xeon/KNL
Secure

ORNL
IBM/NVidia
P9/Volta
Open

LLNL
IBM/NVidia
P9/Volta
Secure

NERSC-9

LBNL
TBD
Open

Crossroads

LANL/SNL
TBD
Secure

Argonne
Intel/Cray TBD
Open

Frontier

ORNL
TBD
Open

El Capitan

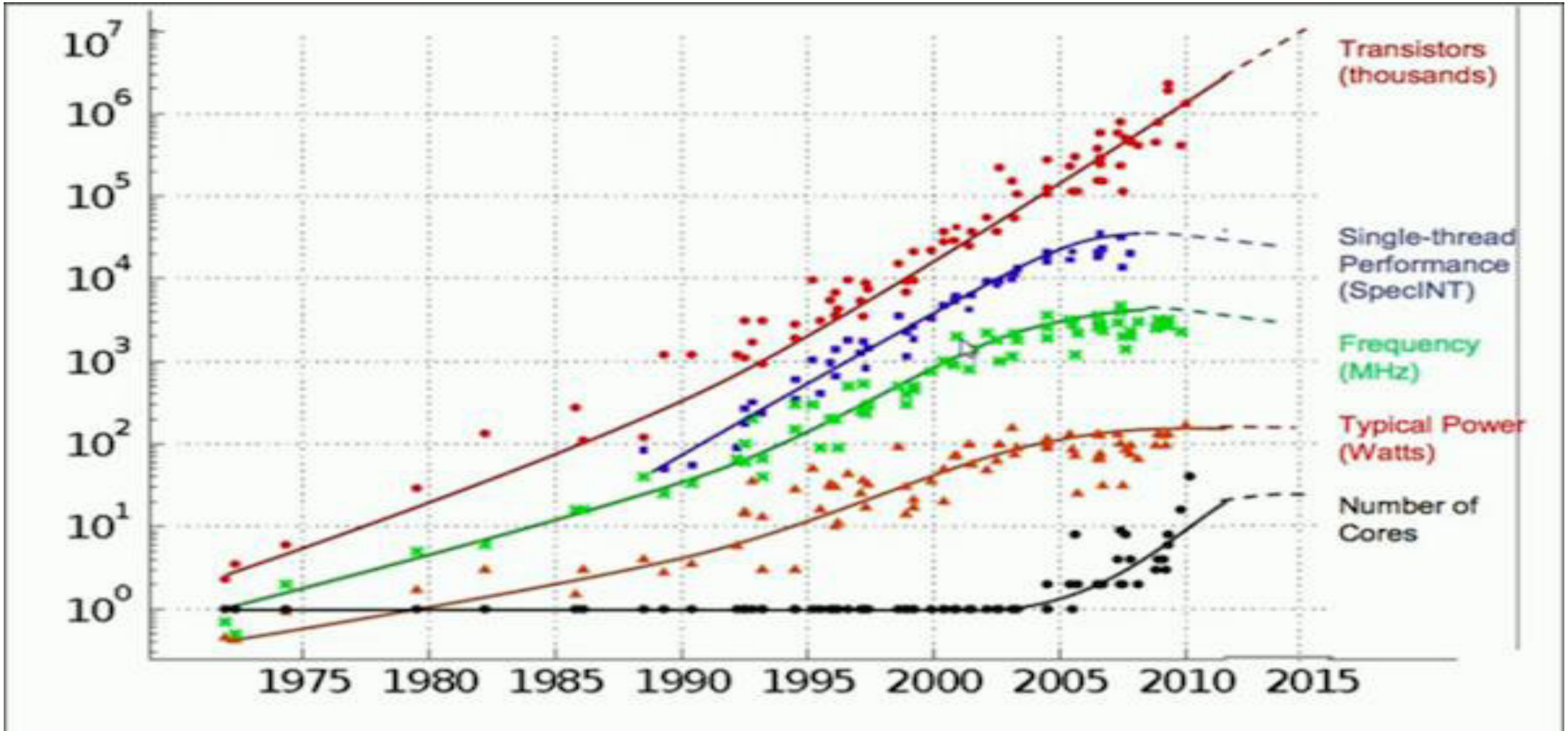
LLNL
TBD
Secure

Building an Exascale Machine

- Why is it difficult?
  - Dramatically improve power efficiency to keep overall power 20-40MW
  - Provide useful FLOPs: algorithms with efficient (local) data movement

- What are the risks?
  - End up with Petscale performance on real applications
  - Exascale on carefully chosen benchmark problems only

Microprocessor Transistors / Clock (1970-2015)



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

Fastest Computers: HPL Benchmark

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
2	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
3	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray Inc.	361,760	19,590.0	25,326.3	2,272
4	Japan Agency for Marine-Earth Science and Technology Japan	**Gyoukou** - ZettaScaler-2.2 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz ExaScaler	19,860,000	19,135.8	28,192.0	1,350
5	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209

EXASCALE COMPUTING PROJECT

Fastest Computers: HPCG Benchmark

Rank	Site	Computer	Cores	HPL Rmax (Pflop/s)	TOP500 Rank	HPCG (Pflop/s)	Fraction of Peak
1	RIKEN Advanced Institute for Computational Science Japan	**K computer** – , SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10.510	10	0.603	5.3%
2	NSCC / Guangzhou China	**Tianhe-2 (MilkyWay-2)** – TH-IVB-FEP Cluster, Intel Xeon 12C 2.2GHz, TH Express 2, Intel Xeon Phi 31S1P 57-core NUDT	3,120,000	33.863	2	0.580	1.1%
3	DOE/NNSA/LANL/SNL USA	**Trinity** – Cray XC40, Intel Xeon E5-2698 v3 300160C 2.3GHz, Aries Cray	979,072	14.137	7	0.546	1.8%
4	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** – Cray XC50, Intel Xeon E5-2690v3 12C 2.6GHz, Cray Aries, NVIDIA Tesla P100 16GB Cray	361,760	19.590	3	0.486	1.9%
5	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** – Sunway MPP, SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93.015	1	0.481	0.4%

Preparing Applications for Exascale

1. What are challenges?

1. What are we doing about it?

Harnessing FLOPS at Exascale

- Will an exascale machine require too much from applications?
  - Extreme parallelism
  - High computational intensity (not getting worse)
  - Sufficient work in presence of low aggregate RAM (5%)
  - Focus on weak scaling only: High machine value of $N_{1/2}$
  - Localized high bandwidth memory
  - Vectorizable with wider vectors
  - Specialized instruction mixes (FMA)
  - Sufficient instruction level parallelism (multiple issue)
  - Amdahl headroom

ECP Approach to ensure useful exascale system for science

- 25 applications projects: each project begins with a mission critical science or engineering *challenge problem*

- The challenge problem represents a capability currently beyond the reach of existing platforms.

- Must demonstrate
  – Ability to execute problem on exascale machine
  – Ability to achieve a specified Figure of Merit

The software cost of Exascale

- What changes are needed
  - To build/run code? *readiness*
  - To make efficient use of hardware? *Figure of Merit*
- Can these be expressed with current programming models?

ECP Applications – Distribution of Programming Models

Node\Internode	Explicit MPI	MPI via Library	PGAS, CHARM++, etc.
MPI	High	High	N/A
OpenMP	High	High	Low
CUDA	Medium	Low	Low
Something else	Low	Low	Low

Bottom Line: All MPI and MPI+OpenMP ubiquitous

Heavy dependence on MPI built into middleware (PetsC, Trilinos, etc)

Will we need new programming models?

- Potentially large software cost + <u>risk</u> to adopting new PM

- However, abstract machine model underlying both MPI and OpenMP have shortcomings, e.g.
  - <u>Locality</u> for OpenMP
  - Cost of <u>synchronization</u> for typical MPI bulk synchronous

- Good news: Standards are evolving aggressively to meet exascale needs

- Concerns remain, though
  - Can we reduce software cost with hierarchical task-based models?
  - Can we retain performance portability?
  - What role do non-traditional accelerators play?

How accelerators affect programmability

- Given performance per watt, specialized accelerators (LOC/TOC combinations) lie clearly on path to exascale
- Accelerators are heavier lift for directive-based language like OpenMP or OpenACC
- Integrating MPI with accelerators (e.g. GPUDirect on Summit)
- Low apparent software cost might be fool’s gold
- What we have seen: Current situation favors applications that follow 90/10 type rule

Programming Model Approaches

- Power void of MPI and OpenMP leading to zoo of new developments in programming models.
  - This is natural and not a bad thing, will likely coalesce at some point

- Plans include MPI+OpenMP but …
  - On node: Many project are experimenting with new approaches that aim at device portability: OCCA, KOKKOS, RAJA, OpenACC, OpenCL, Swift
  - Internode: Some projects are looking beyond MPI+X and adopting new or non-traditional approaches: Legion, UPC++, Global Arrays

Middleware/Solvers

- Many applications depend on MPI implicitly via middleware, eg.
  - Solvers: PetsC, Trilinos, Hypre
  - Frameworks: Chombo (AMR), Meshlib

- Major focus is to ensure project-wide that these developments lead the applications!

Rethinking algorithmic implementations

- Reduced communication/data movement
  - Sparse linear algebra, Linpack, etc.
- Much greater locality awareness
  - Likely must be exposed by programming model
- Much higher cost of global synchronization
  - Favor maxim asynchrony where physics allows
- Value to mixed precision where possible
  - Huge role in AI, harder to pin down for PDEs
- Fault resilience?
  - Likely handled outside of applications

Beyond implementations

- For applications we see hardware realities forcing new thinking beyond implementation of known algorithms
  - Adopting Monte Carlo vs. Deterministic approaches
  - Exchanging on-the-fly recomputation vs. data table lookup (e.g. neutron cross sections)
  - Moving to higher-order methods (e.g. CFD)
  - The use of ensembles vs. time-equilibrated ergodic averaging

Co-design with hardware vendors

- HPC vendors need deep engagement with applications prior to final hardware design

- *Proxy Applications* are a critical vehicle for co-design
  - ECP includes Proxy Apps Project
  - Focus on motif coverage
  - Early work with performance analysis tools and simulators

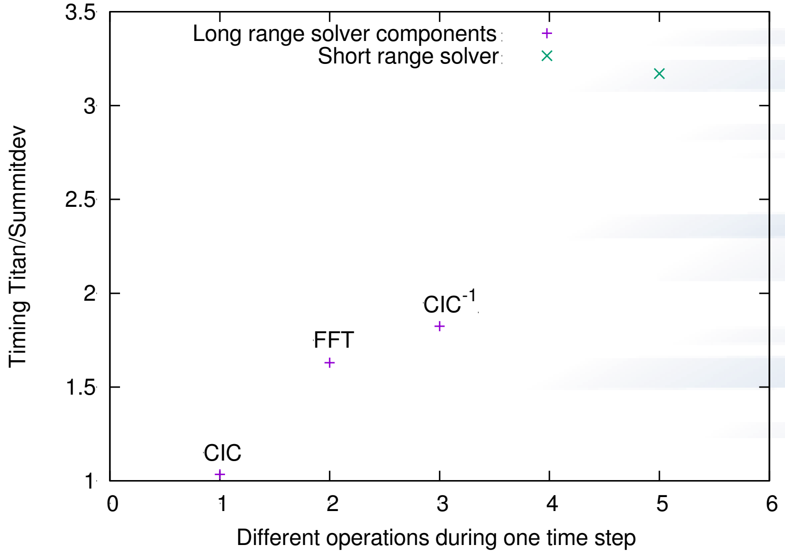
- Interest (in theory) in more complete applications.

First HACC Tests on the OLCF Early-Access System

1.2.1.01 ExaSky
PI: Salman Habib, ANL
Members: ANL, LANL, LBNL

Scope & Objectives

- Computational Cosmology: Modeling, simulation, and prediction for new multi-wavelength sky observations to investigate dark energy, dark matter, neutrino masses, and primordial fluctuations
- Challenge Problem: Meld capabilities of Lagrangian particle-based approaches with Eulerian AMR methods for a unified exascale approach to: 1) characterize dark energy, test general relativity, 2) determine neutrino masses, 3) test theory of inflation, 4) investigate dark matter
- Main drivers: Establish 1) scientific capability for the challenge problem, and 2) full readiness of codes for pre-exascale systems in Years 2 and 3



Speed up of major HACC components on 8 Summitdev nodes vs. 32 Titan nodes (first three points: long range solver, last point: short-range solver).

Impact

- Well prepared for the arrival of Summit in 2018 to carry out impactful HACC simulations
- With CRK-HACC we have developed the first cosmological hydrodynamics code that can run at scale on a GPU-accelerated system
- The development of these new capabilities will have a major impact for upcoming cosmological surveys

Project Accomplishment

- HACC was successfully ported to Summitdev
- The HACC port included migration of the HACC short-range solver from OpenCL to CUDA
- We demonstrated expected performance comparing to Titan and validated the new CUDA version
- We implemented CRK-HACC on Summitdev and carried out a first set of tests

EXASCALE COMPUTING PROJECT

Monte Carlo performance optimization for full core problems

ECP WBS	2.2.2.03: ExaSMR
PI	Steven Hamilton, ORNL
Members	ORNL, ANL, MIT, INL

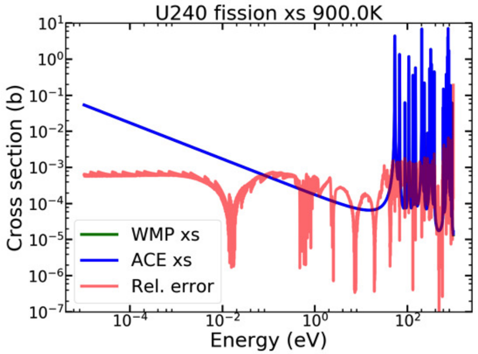
Scope and objectives

- Small Modular Reactor (SMR) Challenge Problems require simulation of very large number of Monte Carlo particle histories to achieve sufficient statistical accuracy
- Current goal is to enhance computational performance based on previous profiling studies
- Additional goal is to improve generation of data libraries for windowed multipole method (WMP)
  - WMP was previously limited to select number of isotopes in nuclear data libraries

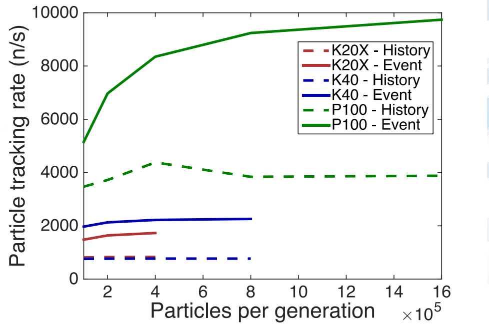
FOM projection for MC transport on Summit.

Machine	Device Type	Device Rate (n/s)	# Devices	Full Machine Rate (n/s)	FOM
Titan	16 core CPU	7.1×10^2	18,688	1.32×10^7	1.0
Summit	1× P100 GPU	9.7×10^3	27,600*	2.68×10^8	20.3

*Based on latest data from olcf.ornl.gov with ∼4600 nodes, 6 GPUs per node



Accuracy of windowed multipole method relative to reference data.



GPU MC performance on depleted fuel benchmark.

Impact

- Improved Monte Carlo particle tracking rate allows reduction in statistical errors
- WMP is now a viable route forward for production Monte Carlo solvers
- Optimization approaches provide insight into optimization strategies for other latency-bound application areas

Project accomplishment

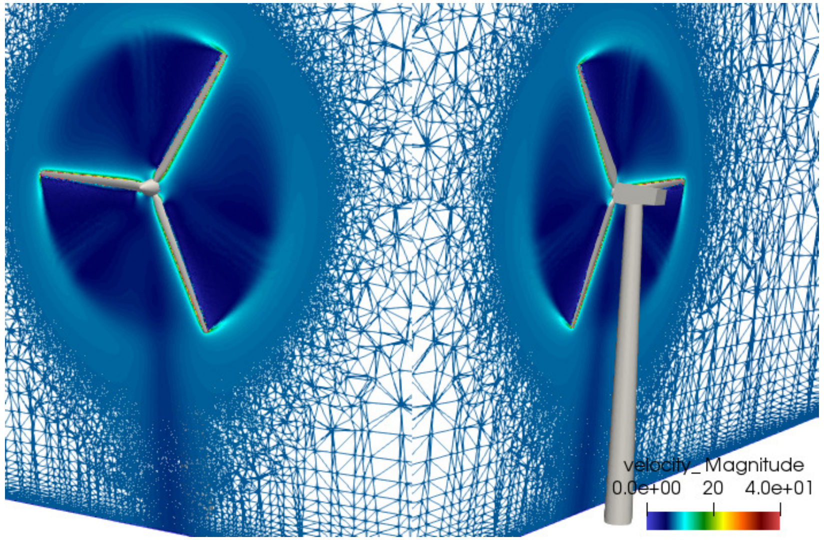
- Realized substantial performance gains on CPU, Intel Xeon Phi, and Nvidia GPU architectures
  - 2-3x speedup across all architectures
- Developed new vector-fitting approach for generation of WMP data libraries
  - Allows processing of data for all nuclides
- Demonstrated KPP figure-of-merit projection of 20 for Summit supercomputer relative to Titan
  - Approximated using previous generation P100 GPU, actual value expected to be larger

FY18-Q1: Deploy production sliding mesh capability with linear solver benchmarking

ECP WBS: 2.2.2.01 ExaWind
PI: Michael Sprague, NREL
Members: NREL, SNL, ORNL, UT

Scope & Objectives

- **ExaWind Objective:** Create a computational fluid and structural dynamics platform for exascale predictive simulations of wind farms
- **Challenge Problem:** Predictive simulation of a wind plant composed of O(100) wind turbines sited over O(100) km^2 with complex terrain
- This milestone is a necessary and critical step in moving towards MW-scale-turbine simulations
  - Establishes baseline performance for a fully resolved sub-MW-scale turbine in an operating configuration



Simulation results for a fully-resolved sub-MW-scale turbine for which the rotor resides in an embedded, rotating "disk" of fluid that is coupled to the surrounding fluid via a sliding-mesh interface. Shown are velocity shadings from the upwind (left) and downwind (right) perspectives.

Impact

- The new sliding mesh capability provides a pathway for efficient simulation of rotating meshes in wind turbine simulations
- Simulating a 1.3B element mesh is a milepost on the pathway to the extreme mesh sizes required for MW-scale-turbine simulations
- Coupling of Nalu with Hypre and MueLu provides insight into, and a comparison platform for, two fundamentally different AMG approaches (classic and smoothed aggregation); highlighted areas for future work

Project Accomplishment

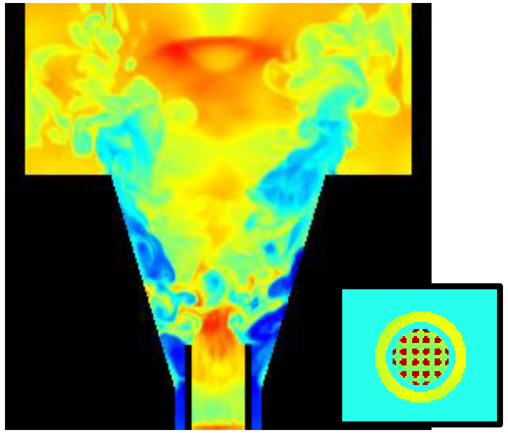
- Deployed and verified a design-order hybrid CVFEM/DG sliding-mesh interface for wind turbine simulations
- Completed transition to Kokkos for interior topology matrix contributions for wind applications
- Coupled the Nalu solver with the Hypre AMG preconditioner and the TIOGA overset library
- Under the ECP ALCC ExaWind allocation on Cori, established baseline timing results for a fully resolved sub-MW-scale turbine
  - Detailed timing breakdown for MueLu/Belos and Hypre solvers
- Successfully simulated sub-MW-scale fully resolved turbine with 1.3B elements

PeleC Embedded Boundary Capability

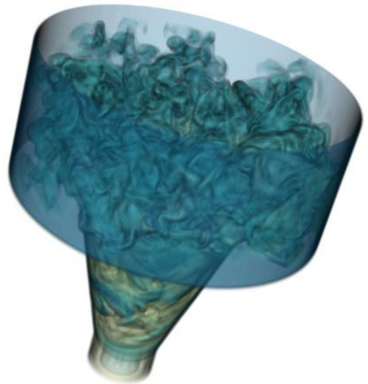
1.2.1.14 Pele
PI: Jackie Chen, SNL
Ray Grout, NREL; Jon Rood, NREL;
John Bell, Marc Day, Dan Graves LBNL

Scope & Objectives

- The goal of this project is to provide a simulation capability for first-principles (DNS) and near-first principles (DNS/LES hybrids) simulations of turbulence-chemistry interactions in conditions relevant to practical combustion devices, including turbulence, mixing, spray vaporization, low-temperature ignition, and flame propagation.



Z-momentum on cutting plane through center of combustor geometry (body cells not blanked, inlet velocity through central pipe in inset).



Volume rendering of density field matching image at left

Impact

- Accurate simulation of combustion at high pressure such as conditions in a diesel engine requires modeling non-ideal fluid behavior, particularly for large hydrocarbons
- Four year demonstration problem is a single sector of a gas turbine combustion; the geometry of the flame holder is needs to be captured to generate recirculation zones that anchor the flame.

Project Accomplishment and Next Steps

- Cartesian cut cell implementation in PeleC allows simulation of complex geometry using explicit diffusion treatment and method of lines approach to hyperbolic treatment
- Capability demonstration is ~30x faster than start of project baseline and 5x slower than proof of concept created by AMReX and tailored for gamma-law gas dynamics
- Calculation of diffusive and advective fluxes needs to be coordinated to improve computational throughput and reduce memory usage
- Performance engineering of initial code for more general cases (multivalued, vector potential) is next major step

Summary

- Major challenge for mission-critical HPC applications to get proportional performance moving toward exascale

- From application perspective high risk in being passive
  - Engage now with HPC vendors
  - Be aware of emerging technologies, particularly new ideas for programmability
  - Drive new science/engineering opportunities and numerical approaches by key features of hardware