High Performance Computing for Dark Energy Missions

Julian Borrill
Computational Cosmology Center,
Lawrence Berkeley National Laboratory

Perspectives

- HPC user and user representative
  - Most computationally challenging CMB data analyses
  - US (DOE,NSF,NASA) & Europe (CINECA,CSC,BSC)
  - Representing HEP on NUGEX
- HPC systems tester and reviewer
  - NERSC procurement benchmark, early user role
  - NASA HPC procurement reviews
- US Planck Computational Systems Architect
  - Ensuring appropriate resources & systems available for US (and European) Planck members
- Long-term DES Data Management reviewer

Data Processing Elements

- Data reduction
  - Single pipeline generating official data products
  - Some time-critical elements (eg. supernovae)
  - Largely embarrassingly parallel by observation or field
- Data analysis
  - Multiple pipelines extracting a range of science from reduced data
  - Multiple versions of many pipelines
  - Significant fully parallel elements (eg. simulations)
- Data distribution
  - Within collaboration
  - To the world

Computing Needs (I)

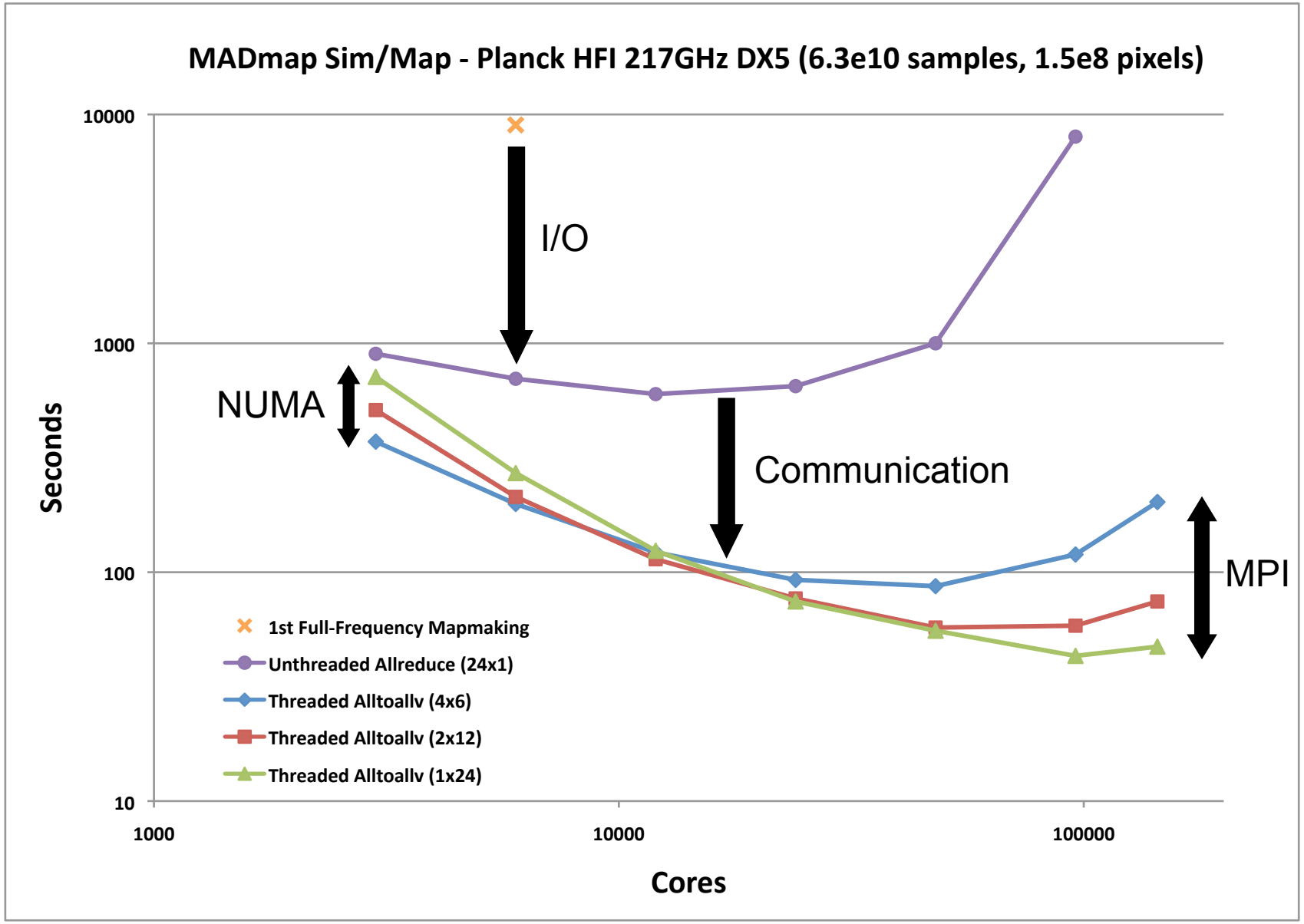
- Resources
  - Cycles – GPU, Many Core (per node, total)
  - Memory – Heterogeneity & deeper hierarchy
  - Fast storage – Flash
  - Archival storage – Cloud
  - Network – OpenFlow

- Both needs and resources evolve over mission lifetime.
- Resource evolution is driven by other markets & agencies.
- Transformative changes are coming.
- Resources are only available as part of a whole system.

Resource Hierarchy

- Data movement is everything, cycles are free.
- Optimize to the hierarchy of subsystem costs (~time):
  - network: localize data
  - IO: re-calculation over write/read
  - inter-process communication: hybridize (MPI+threads)
  - on-node communication:
    - threading overhead, GPU bus challenge
  - memory hierarchy:
    - NUMA effects, power challenge & dark silicon
  - cycles: trade for all the above
- Cost hierarchy exacerbated by scale, sharing
- Cost dependence on system & scale imply auto-tuning

Example – Monte Carlo SimMap



Computing Needs (II)

- Capacity
  - very many jobs each using a small number of nodes
- Capability
  - a small number of jobs each using very many nodes
- Collaboration
  - account availability, capacity, security
- Continuity
  - system evolution/growth over mission lifetime
- Control
  - scheduling, software environment
- Cost
  - hardware + infrastructure + administration

Resources

- Local/Dedicated Cluster

  Pro: Capacity*, Control

  Con: Capability, Collaboration, Cost, Continuity

- Grid/Cloud computing

  Pro: Capacity, Collaboration, Continuity*, Cost*

  Con: Capability, Control

- @home computing

  Pro: Capacity, Cost, Continuity

  Con: Capability, Control, Collaboration

- Supercomputer

  Pro: Capacity, Capability, Cost, Collaboration, Continuity

  Con: Control*

Resource Providers

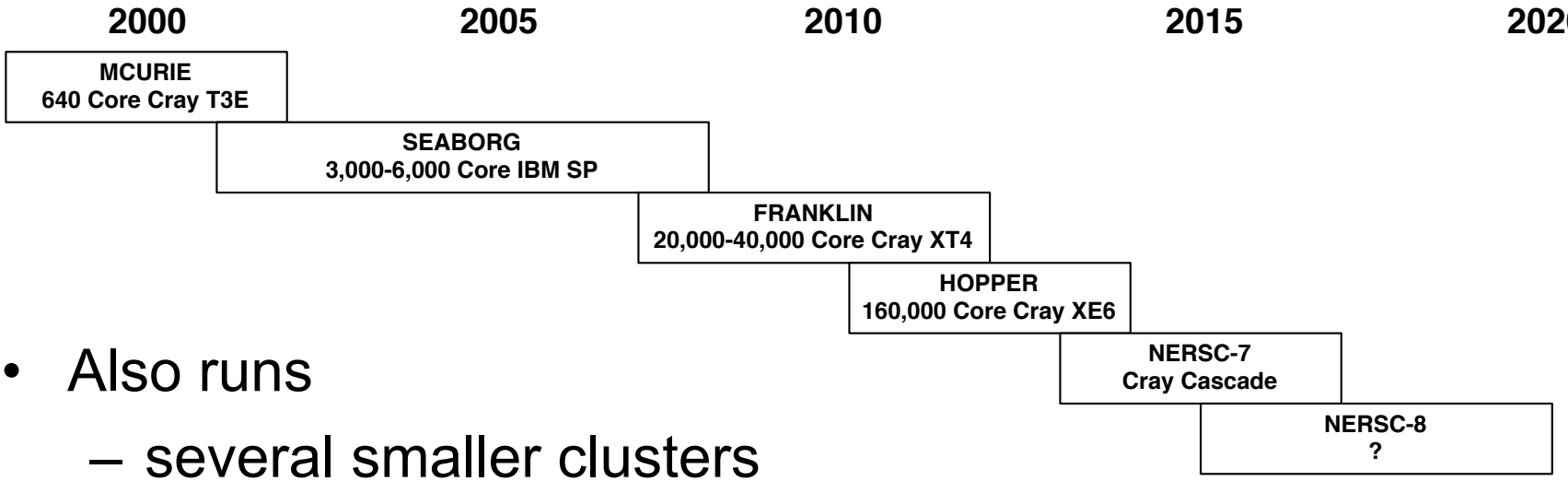
- Local clusters – universities, federal laboratories
- Dedicated clusters – space & larger suborbital missions
- Grid computing – NSF (XSEDE), Europe (PRACE)
- Cloud computing – Amazon etc
- @home computing – www
- Supercomputing (top 20 systems)
  - Europe (CINECA, CEA)
  - NASA Ames (Pleiades)
  - NSF NCSA (Blue Waters)
    - SDSC (Gordon)
  - DOE NERSC (Hopper, NERSC-7), LCFs (Mira, Titan)
    - NNSA (Sequoia, Cielo)

The Planck Example

- Dedicated clusters at two European data processing centers (Paris/HFI, Trieste/LFI).
- Each performs official pipeline reduction & analysis of one instrument's data.
- Problems:
  - very limited resources
    - constrain official analysis
    - preclude other analyses/analysts
  - very restrictive pipeline designs
    - locked to one instrument's data
    - locked to DPC-specific infrastructure
- US goal – provide resources to support fully collaborative analyses of both data sets, both individually and together.

NERSC

- DOE's general-purpose supercomputing center.
- Runs 2 high-end systems simultaneously.
- Deploys a new top-10 system every ~2 years



- Also runs
  - several smaller clusters
  - large global file-system
  - modular software stack
  - project disk spaces

US Planck At NERSC

- Planck use evolved from previous CMB work at NERSC
  - substantial annual allocation of resources
  - open access for all Planck data analysts
  - excellent support, eg. priority boost during critical times
- Formalized by 2007 NASA/DOE MoU
  - guaranteed minimum allocation throughout mission
    - up to 10M CPU-hours/year
  - 100TB unpurged space on global filesystem
  - locate dedicated US Planck resources at NERSC
    i. 256 core stand-alone cluster (2008-11)
    ii. 640 dedicated system cores & queue (2012-15)
    - additional 2 – 5M CPU-hours/year

Additional US Planck Resources

- IPAC cluster(s)
  - US entry point for Planck data
  - source of Early Release Compact Source Catalogue (formal US deliverable)
  - home of US Planck archive

- planck@home
  - distributing/collating most embarrassingly parallel cosmological parameter calculations

Conclusions

Understand your data challenge:

- Know the scaling and efficiency of your algorithm and its implementation, both in theory and practice.
- Make *informed* algorithm/science trade-offs
  - often implementation is the issue
  - Moore's Law is your friend!
- Remember that the computational challenge is dynamic
  - implementations evolve with the scale and balance of each new generation/class of HPC system.
- Don't shoot yourself in the foot!
  - build in data efficiency from the outset.
- Find the resources for the problem, not the problem for the resources.

Conclusions

- Match the computing need to the resource
  - most time-critical: dedicated clusters
  - most embarrassingly parallel: @home
  - most computationally challenging: supercomputers
  - widest collaboratory: supercomputer centers

- Exploit the additional opportunities
  - dedicated clusters: sponsorship
  - @home: public outreach
  - supercomputing: inter-agency, trans-disciplinary & industrial partnerships