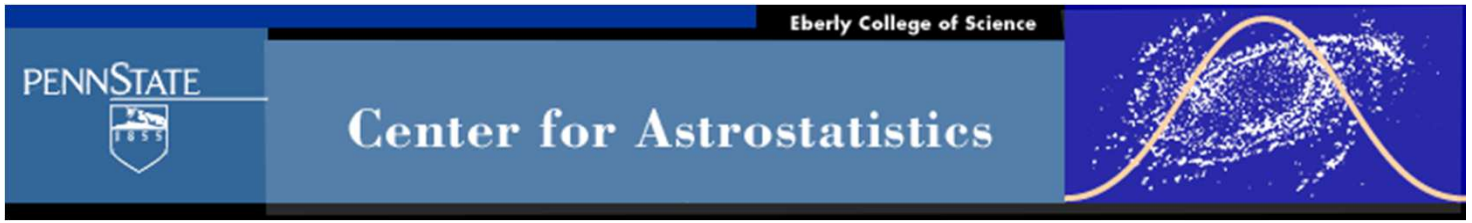
Motivation for High-Performance Computing for Bayesian Analysis of Astronomical Data Sets

Eric B. Ford (Penn State)

Bayesian Computing for Astronomical Data Analysis

June 11, 2014



BayesComp Instructors

Lecturers

- Eric Ford (Penn State, Astronomy)
- Pierre-Yves Taunay (Penn State, Research Computing Center)
- Tamas Budavari (JHU, Physics & Astronomy)
- Jessi Cisewski (CMU, Statistics)
- Daniel Lee (Columbia, Statistics)
- Daniel Foreman-Mackey (NYU, Astronomy)
- Tom Loredo (Cornell, Astronomy)

Lab Assistants:

- Robert Morehead, Benjamin Nelson, Megan Shabram (Penn State Graduate Students)

Advantages of Bayesian Approach

- Rigorous statistical basis for performing inference
  - Parameter estimation
  - Model comparison
  - Prediction
- Explicitly account for
  - Prior information (e.g., physical constraints, previous work),
  - Correlated parameters, and
  - Non-Gaussian uncertainties
- Sounds great in theory, but…

Bayesian Inference Requires Computing Lots of Integrals

- Old-School Bayesian “Solution”
  - Used simple models and conjugate distributions
- But we often have physical models that:
  - May not permit using conjugate distributions
  - Are highly non-linear
  - Are time consuming to evaluate
  - Have many model parameters
  - Can be under-constrained by available data

Bayesian Inference Requires Computing Lots of Integrals

- Modern Bayesian Approach
  - Apply Modern Computing Power & Clever Algorithms
- Many Bayesian analyses are practical using
  - Workstation
  - A Few Basic Algorithms:
    - Quadrature Integration
    - Monte Carlo Integration
    - Markov chain Monte Carlo
    - Importance Sampling

Rise of Bayesian Inference

- Bayesian inference has become increasingly common, largely thanks to:
  - Rapid increase in computational power
  - Application of Markov chain Monte Carlo (MCMC)

But Some Problems are Hard

- Quadrature Integration:
  - Breaks down for large dimensions
- Monte Carlo Integration
  - Requires unrealistic number of evaluations
- Standard Markov chain Monte Carlo
  - Can converges too slowly if many parameters, significant correlations, or multi-modal posterior
- Importance Sampling
  - Can be difficult to implement for complex problems (e.g., finding good importance sampling density)

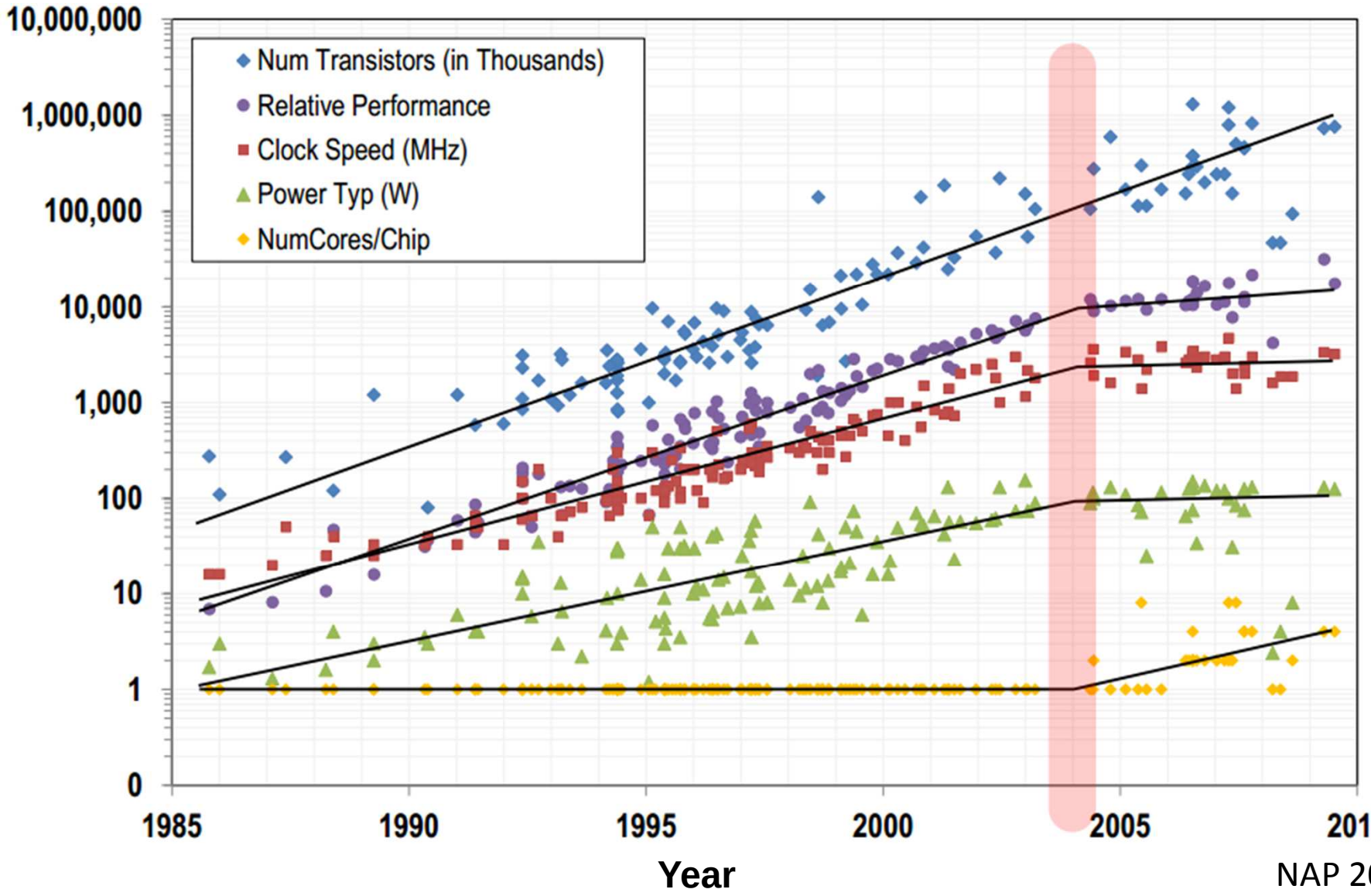
In Near Future, Greater Fraction of Problems Will be “Hard”

- Increasingly large data sets
  - Particularly surveys like SDSS, HETDEX, LSST
- Need to combine multiple types of observations
  - E.g., SuperNovae, Large scale structure
- Increasing complexity of models to compare with data

Two Basic Solutions

- Increase available computational power
  - Buy/apply faster hardware
  - Increase level of parallelism
- Develop/apply more efficient algorithms
  - Astronomers can learn from statisticians

Evolution of Computing Power



NAP 2011

Two Basic Solutions

- Increase available computational power
  - Buy/apply faster hardware
    - Computers aren't getting much faster anymore
  - Increase level of parallelism
    - Topic of Wednesday lessons/labs
- Apply more efficient algorithms
  - Astronomers can learn from statistians
    - Topic of Thursday & Friday lessons/labs

Goals for This School

Wednesday

- Which architectures/programming tools are likely to result in efficient parallelization of a given problem?

Thursday & Friday

- What algorithms are available that are likely to result in more efficient convergence for a given problem?

- Not time to become expert at any, but enough to help you choose which techniques to pursue in detail

Goals for Today

Which architectures/programming tools are likely to result in efficient parallelization of a given problem?

- Basic properties of different computer architectures
  - Modern CPU, Cluster, GPU
- Parallel programming toolboxes
  - Fast functions: OpenMP (e.g., parallel for loops within one workstation)
  - More involved functions: MPI & map-reduce (e.g., clusters, cloud)
  - Very large number of function evaluations: GPUs