Bioinformatics Compute Requirements

Marshall Peterson, VP

mrpiii@attglobal.net

Accelerating Biology

Goal

Accelerating biology through intelligent application of innovative HPTC Technologies

What is Needed

A flexible, scalable, secure, highly performing, highly available computing infrastructure that adapts to a wide range of continuously evolving and challenging demands.

The Past as Prologue – Celera Example

Celera Foundation

- Craig Venter
- Gene Myers
- Granger Sutton
- Mark Adams
- Vivien Bonazzi
- Ham Smith
- Paul Gilman
- TIGR

What Celera Anticipated

- Assembly
  - 1 Terabyte memory
  - 6-8 Alpha processors
  - 6 – 8 Months
  - 25 MB
- GCMP
  - Discrete Systems
  - 100 processors
  - 76 MB
- Target -- September 2002
- 90,000 to 120,000 genes
- 4-6B base pairs
- Data Ownership
- $26M

What Transpired

- Assembly
  - 32 GB
  - Two separate Assemblers (two approaches)
  - 100+ loosely coupled CPUs
  - 16,000 CPU hours
  - 20+ Assemblies
- GCMP
  - Shared Systems
  - 800 processors
  - 100+ TB
- Target → May 2000 (9/8/99)
- 30,000 genes
- 3.12B base pairs
- Business Model
- Data Explosion - Annotation

The Bottom Line

Biology, specifically the nature of disease, is much, much more complex than originally anticipated. It is as much a data challenge as it is a compute challenge.

- Government
- Industry
- Small Companies

Commercial Bioinformatic Infrastructure Challenges

- Flexibility
- Implementation Speed
- Extremely High File System Throughput
- Ease of Use
- Scalability of Capacity
- Technology Refresh
- Distributed Data
- Distributed Computes (?)
- People

HPTC – A New Paradigm

Commercial HPTC is a Significant Extension to Traditional HPTC Capabilities

- Traditional -- not usually a time critical component
  - Spooks
  - Nukes
  - Aerospace
  - Weather
  - Automotive
- Time critical role Life Sciences businesses
  - HPTC is an integral component of the production pipeline

Scalability of Capacity

Today's *capacity* is tomorrow's *incapacity*….

Compute capacity, storage capacity, system throughput and supporting infrastructure must be flexible, scalable and be designed to accommodate the seamless integration of new and upgraded technologies that will be required to meet the demands of computational biology

Flexibility

- Industry tools are in rapid development
  - Limited ability to predict resource requirements
  - Scaling dramatically
- Opportunities are limited by compute capacity
  - Venter's Law
  - High-Throughput Docking
  - Structure Determination
  - Finding Small Molecule Targets
- Development in "production" is a business reality
- Wide variability of application profiles (BLAST and Assembly)
  - Computes (Scalar, Parallel, Large Memory, Small Memory)
  - I/O (Stream, Random, Massive Datasets, Search)

Implementation Speed

- Business Need is Immediate
  - Sept. ’02 -- Scheduled
  - May ’00 -- Actual
- Solutions
  - Complex
  - Multi-Vendor
  - Large-Scale
  - Dynamic
- Commercial
  - This is Not just about Computers

System Bandwidth

- Speed Matters
- Search
- Stream
- Any file, any CPU
- Bottleneck
- Unknown relationships
- Disparate data sources
- Disparate locations
- Ownership

Ease of Use

Solution …. Solution…. Solution…

- Management Tools
  - Manage, Measure, Monitor
- Documentation
- Life Cycle Support
  - Technology Renewal
- Integration with existing operational environments
  - New Technology Integration
  - CA
  - Netcool
  - Tivoli
- Application Development
  - MPI?

Technology Refresh

The day you deploy new technology you must have a plan to replace it.

- Solutions will accelerate technology introduction and innovation
- Refresh into Production
- As much a financial issue as a technology issue

People

- Limited
- Limited
- Limited…
- Expensive
- Non-existent

Future Challenges

Speed {still} Matters

How Technology Must provide Solutions
The DOE Opportunity

The Challenge

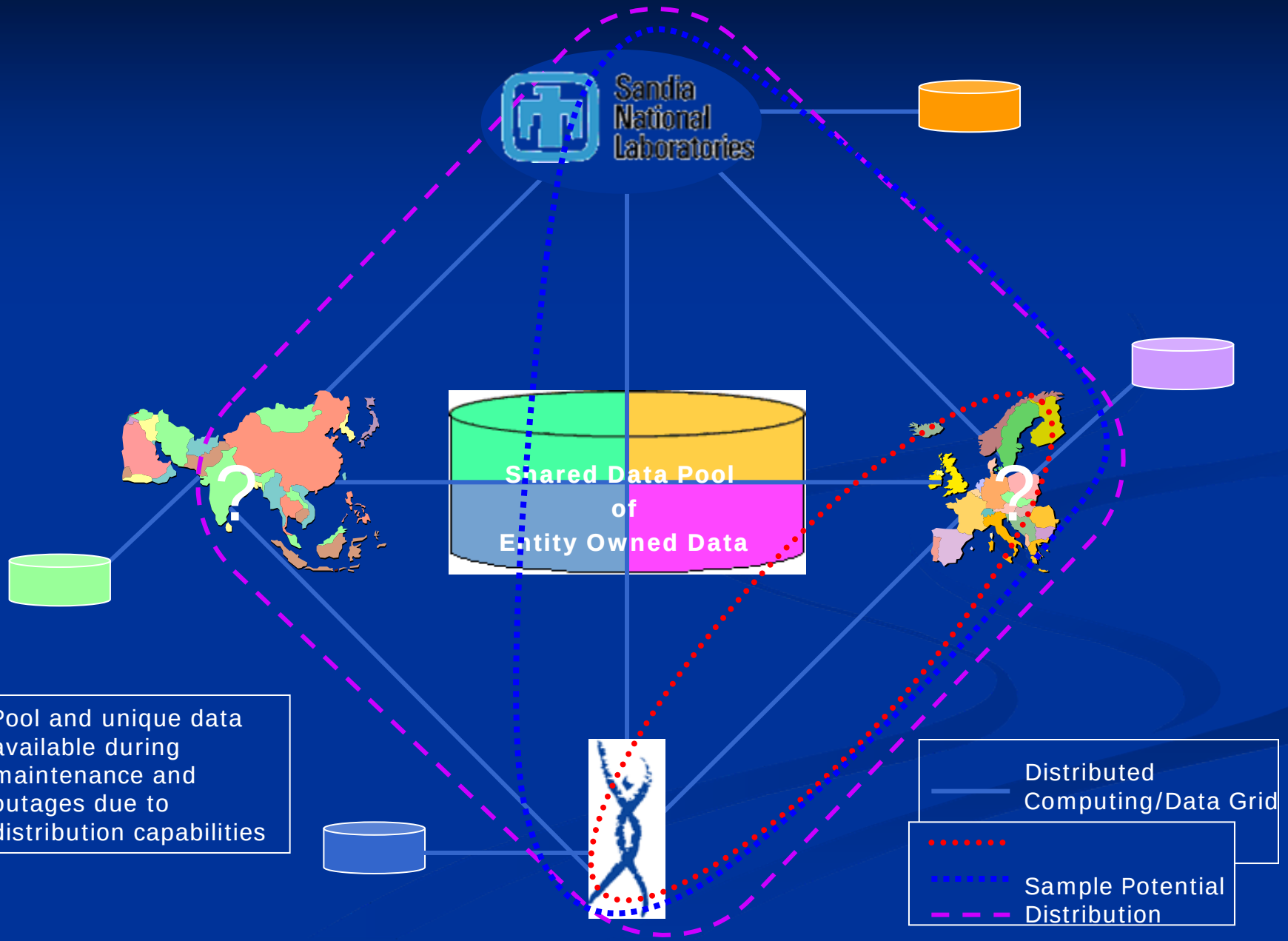
An *integrated computational capability* consisting of a low and high latency massively parallel hardware architecture, scalable to multiple petaflops with the companion software suite and distributed resource management essential to providing the US what **MUST** be the world's fastest and most cost-effective distributed computing environment.

The New Paradigm – Continued

We need a **revolution**…

…Not just an evolution

Distributed Scalable Computing



Distributed Scalable Computing, continued….