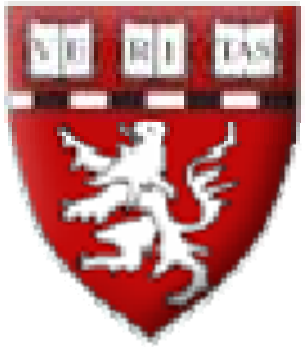
HPC for Biomed Applications The Harvard Orchestra Experience

Marcos Athanasoulis, Dr.PH
Director, Information Technology
Harvard Medical School



Outline

- About HMS
- Context
- Why Shared Biomed HPC is different
- About Orchestra today and tomorrow
- Why it matters
- Biomed HPC 2010

About the Longwood Medical Area

- 213 Acres, 37,000 employees, 15,000 students
- 21 institutions
- 2.15 million in- and outpatient visits
- Forty-seven percent of all hospital-based outpatient clinical visits,
  and fifty-one percent of all inpatient admissions in Boston
- Forty-seven percent of all staffed beds in Boston
- 15,016 births in the LMA

HMS Affiliated Research – Longwood

- Four of the top five Independent Hospital recipients of NIH funding nationwide
- Massachusetts was the number two state recipient of National Institutes of Health (NIH) funding
- Boston is ranked as the number one city in the nation for NIH support
- If the LMA were ranked as a city, it would be number three for funding, after New York and before Philadelphia. If the LMA were ranked as a state, it would be number eight, after North Carolina, and before Washington.
- National Institutes of Health (NIH) awards more than doubled for the LMA institutions from $302 million to $722 million over the decade between FY 1991 and FY 2001

What makes Biomed HPC Different?

- Larger problem space
  - Whole genome processing
  - Whole ‘Ome processing
  - Image Processing
  - Simulations
  - Matlab, R, SAS, etc
  - Everything Else
- Bursty Usage
  - Processing power is not always the bottleneck
  - Most work is “embarrassingly parallel”

Biomed HPC Differences (cont.)

- Researchers
  - Funding challenges
  - Grant funding limitations and requirements
  - Everyone is a CIO
- Systems Diversity
  - Plethora of small clusters
  - General lack of centralization
  - White boxes to blue genes
- Lack of IT training

About HPC @ HMS

- Today: Orchestra Phase I
  - Modest shared cluster
  - 1000 processor cores
  - 600TB attached NAS storage
  - Interconnect: Gigabit Ethernet
  - Subsidized user contribution model
  - BUT, MOST computing happens under the desk and behind the curtain!

About HPC @ HMS (cont.)

- This Fall: Orchestra Phase II
  - $3.7 Million NIH S10 ARRA Award
    - Will support 30 existing NIH researchers
  - 6,000K processor cores
  - Petabyte of storage
  - 10g Ethernet and Infiniband
  - High Performance Scratch Space
  - Small, integrated GPU cluster
  - Modest SMP options (24 cores)

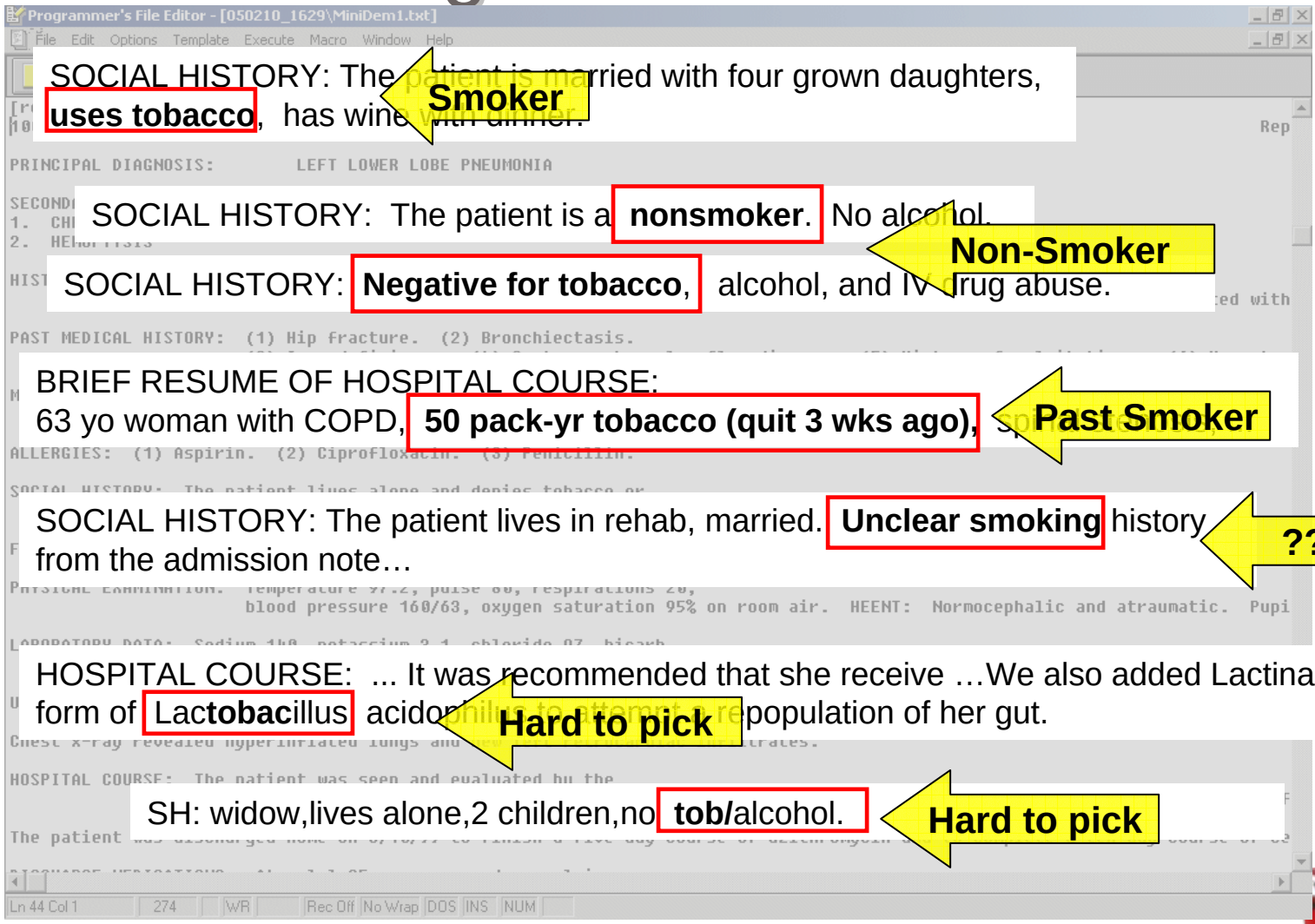
Shared Capacity @ HMS

- Critical to the success of the shared infrastructure model
  - ½ the nodes have been contributed because of shared usage models
  - Contributors get guaranteed access to their nodes but others can use them when they were idle
  - NOTE: This approach would fail in many other disciplines
- LSF used for queue management
  - Time limited queues
  - All and shared queues

Storage Needs and Architecture

- 70% annual growth in storage with an accelerating growth rate
  - Driven largely by next generation sequence analysis and high resolution microscopy
- Implemented Isilon Storage Arrays
  - 2 600TB clusters mirrored in different facilities
  - 6 weeks of checkpoints (daily for last week)
  - Uses NFS

Challenge: Natural Language Processing



SOCIAL HISTORY: The patient is married with four grown daughters, **uses tobacco**, has wine with dinner.

Smoker

SOCIAL HISTORY: The patient is a **nonsmoker**. No alcohol.

SOCIAL HISTORY: **Negative for tobacco**, alcohol, and IV drug abuse.

Non-Smoker

BRIEF RESUME OF HOSPITAL COURSE:
63 yo woman with COPD, **50 pack-yr tobacco (quit 3 wks ago),**

Past Smoker

SOCIAL HISTORY: The patient lives in rehab, married. **Unclear smoking** history from the admission note…

???

HOSPITAL COURSE: ... It was recommended that she receive …We also added Lactinax, oral form of Lac**tobac**illus acidophilus to attempt a repopulation of her gut.

Hard to pick

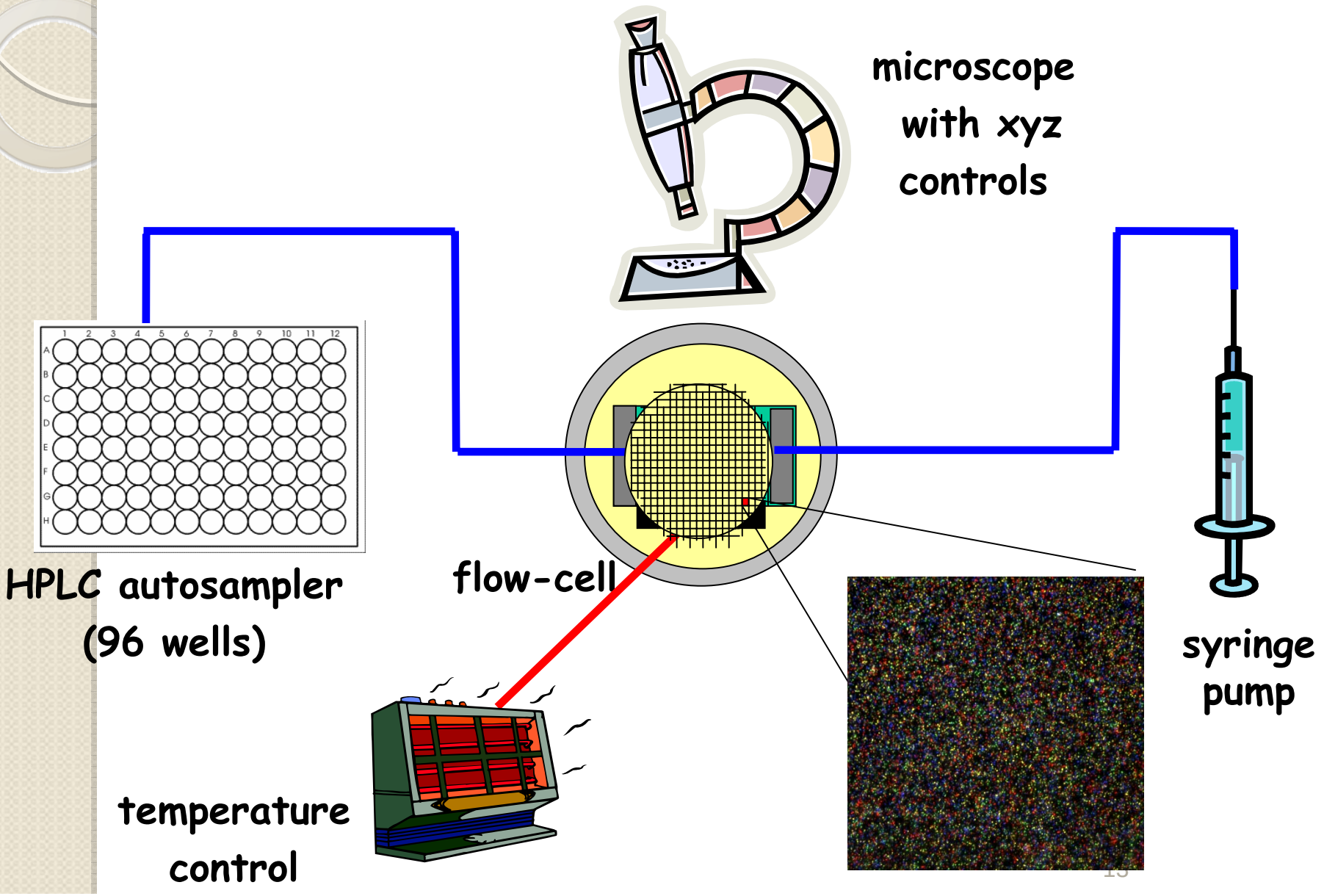
SH: widow,lives alone,2 children,no **tob/**alcohol.

Hard to pick

Challenge: Whole Omes

- Current cost 100K
- Working on <$1,000 whole genome
- High Throughput Instrumentation
  - $250-$500 for 500,000 SNP's
  - $50-100K for good quality phenotyping of 100K++ individuals
  - What about the samples (consented)
    - $650/patient
    - Dozens a week
    - Wait in clinic: $450+/patient

Sequencing Equipment

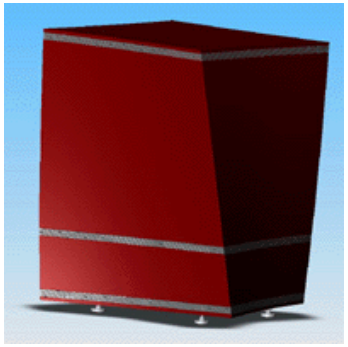


2nd-generation sequencing

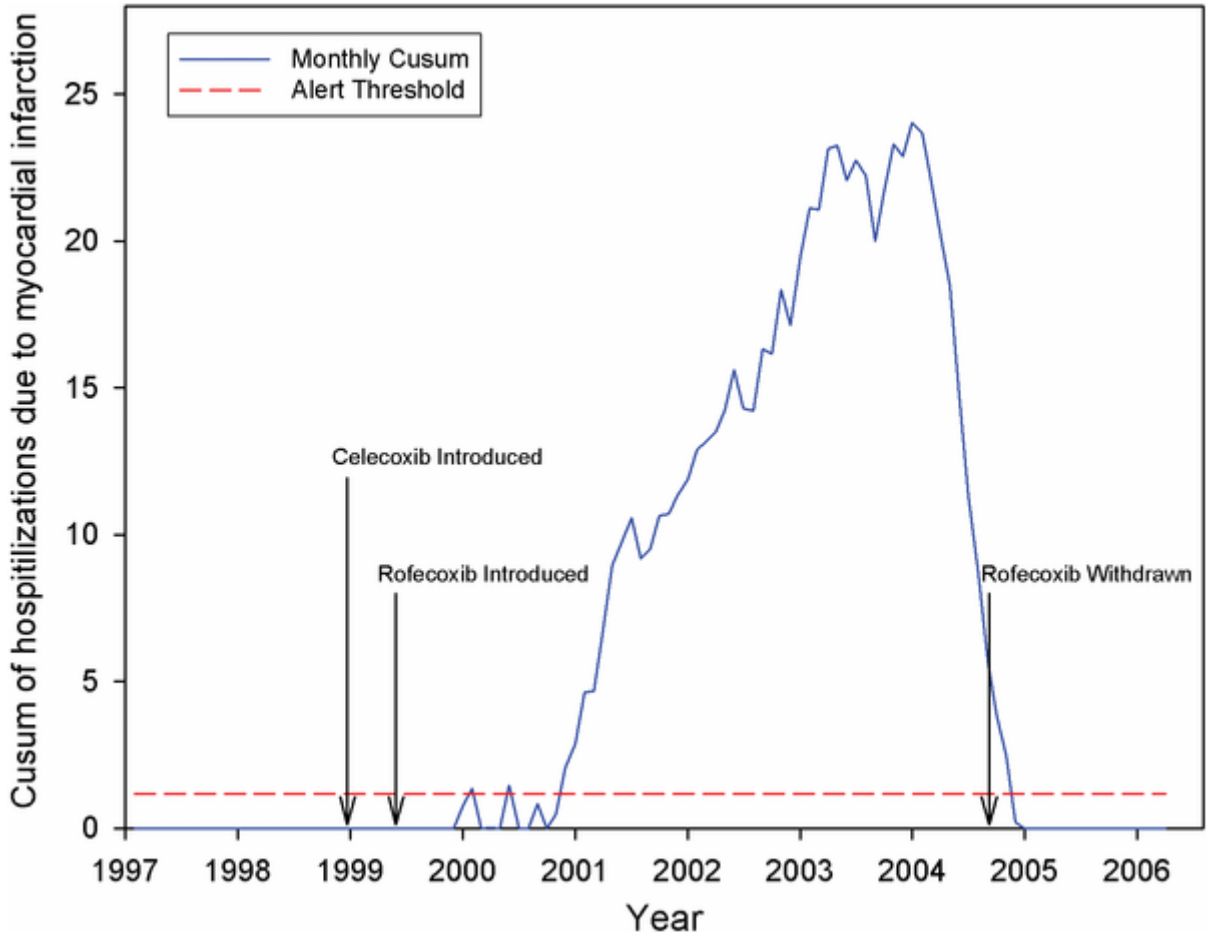
Harvard-model-F07: **$106K** incl. computer. $14K support.
Open-source software, hardware, wetware Reduce reagent volume & per vol cost 100X each.

E07 (Nikon)

F07



Challenge: Everything to Everything



Biomed HPC Leadership Summit

- 150 leaders in biomedical HPC
- Networking, birds of a feather and user polls
- 2010 Summit to convene October 17th – October 19th in Boston MA
- http://biomedhpc.med.harvard.edu

Thank you

- Questions, comments:
  - marcos@hms.harvard.edu