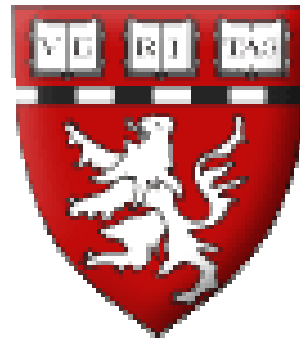




HPC for Biomed Applications

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



HPCS | Cyprus | 2008

Outline

- About HMS
- Why Biomed HPC is different
- Context
- Results from Biomed HPC 2007 Summit
- Predictions and recommendations



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
 - 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



About HPC @ HMS

- Today:
 - Modest shared cluster
 - ~1000 processor cores
 - 100TB 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.)

- **Tomorrow:**
 - Mid-scale cluster and Harvard Grid
 - 10-20K processor cores
 - Petabyte of storage
 - Parallel file system
 - 10g Ethernet or Infiniband
 - More centralized



Open Source @ HMS

- Majority of computational work done using open source or homegrown tools
- Transition from commercial to open source infrastructure software (e.g. Oracle → MySQL)
- Challenges in creating the grid
 - Tyranny of the immediate
 - HIPAA and regulatory barriers
 - Data transfer and environment challenges



Challenge: Natural Language Processing

Programmer's File Editor - [050210_1629\MiniDem1.txt]

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

PRINCIPAL DIAGNOSIS: LEFT LOWER LOBE PNEUMONIA

SECONDARY: SOCIAL HISTORY: The patient is a **nonsmoker**. No alcohol. **Non-Smoker**

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

PAST MEDICAL HISTORY: (1) Hip fracture. (2) Bronchiectasis.

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

ALLERGIES: (1) Aspirin. (2) Ciprofloxacin. (3) Penicillin.

SOCIAL HISTORY: The patient lives alone and denies tobacco or alcohol use. **Unclear smoking history** **???**

PHYSICAL EXAMINATION: Temperature 97.2, pulse 88, respirations 20, blood pressure 160/63, oxygen saturation 95% on room air. HEENT: Normocephalic and atraumatic. Pupils equal and reactive to light.

LABORATORY DATA: Sodium 148, potassium 3.4, chloride 97, bicarbonate 24, glucose 108, creatinine 1.2, urea nitrogen 18, hemoglobin 12.5, hematocrit 38, white blood cell count 12,000, differential 78% neutrophils, 18% lymphocytes, 4% monocytes, 0% eosinophils, 0% basophils.

HOSPITAL COURSE: ... It was recommended that she receive ... We also added Lactobacillus acidophilus to attempt repopulation of her gut. **Hard to pick**

HOSPITAL COURSE: The patient was seen and evaluated by the physician on 07/19/12. She was found to have a left lower lobe pneumonia. She was started on amoxicillin-clavulanate and azithromycin. She was also given Lactobacillus acidophilus. She was discharged home on 07/19/12. She was advised to continue with her current medications and to return to the clinic for follow-up. **SH: widow, lives alone, 2 children, no tob/alcohol.** **Hard to pick**

Ln 44 Col 1 | 274 | WR | Rec Off No Wrap DOS INS NUM

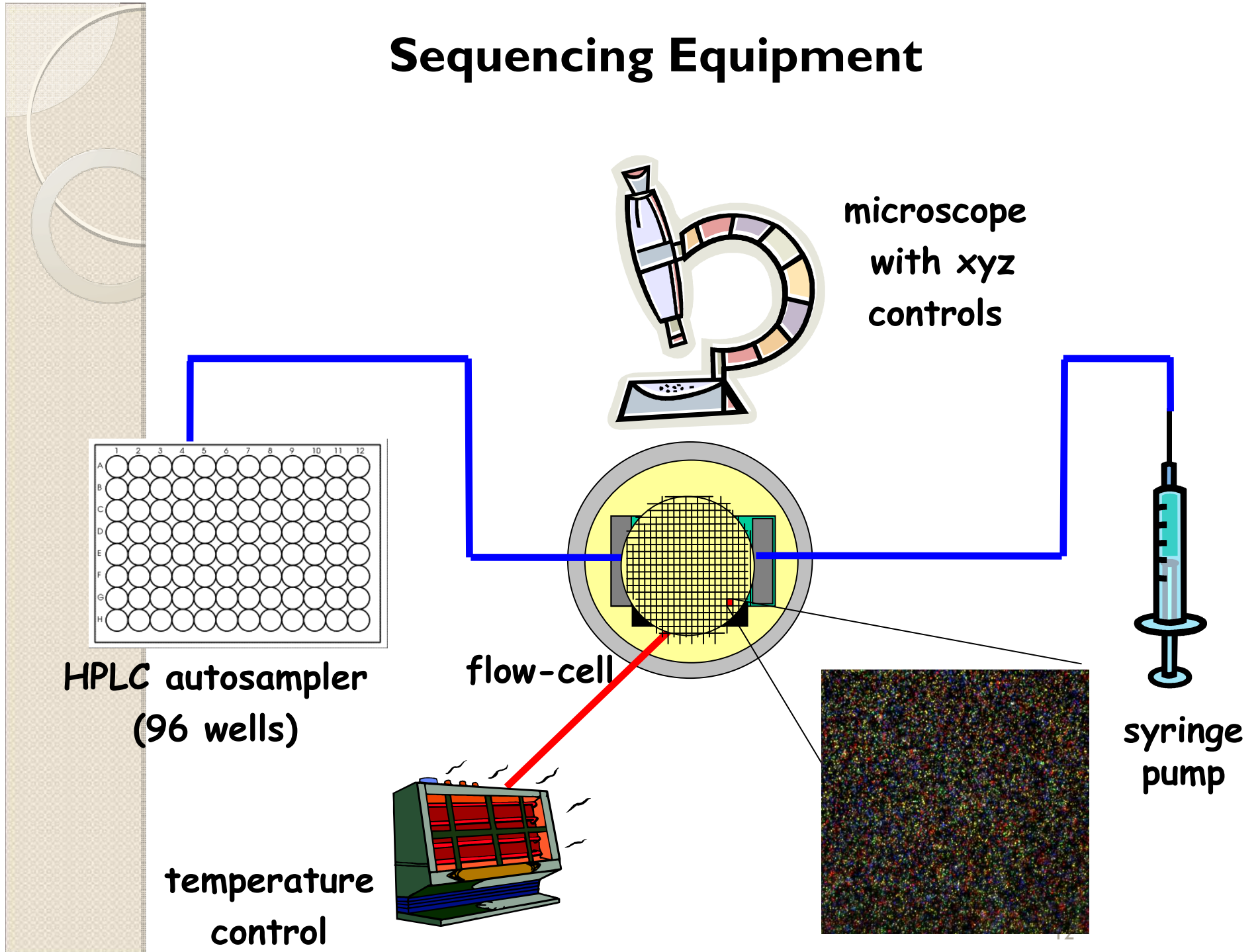


Challenge: Whole Omes

- 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)

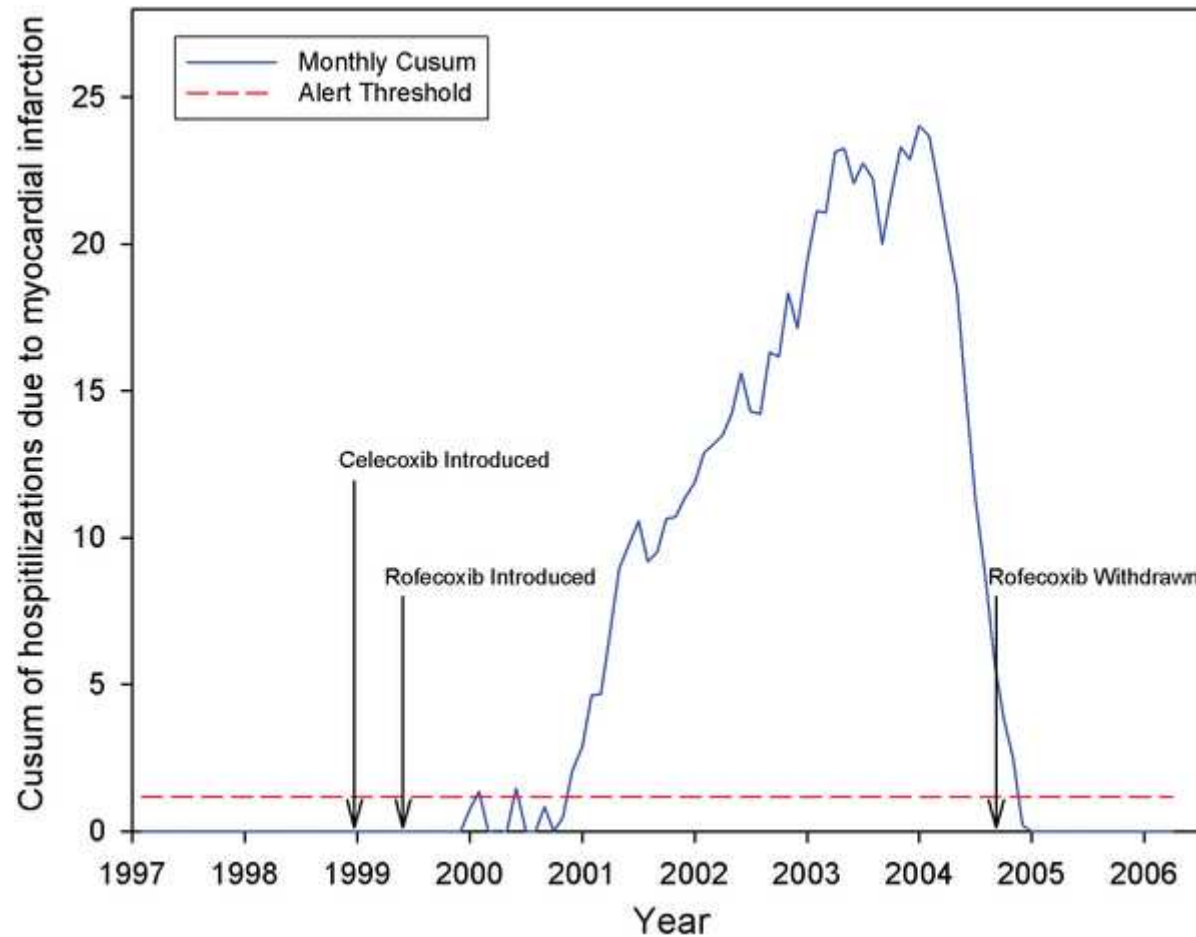


F07





Challenge: Everything to Everything



Biomed HPC Leadership Summit

- 200 leaders in biomedical HPC
- Invitation only
- For the people who build and implement HPC and not for the users of HPC
- 2008 Summit to convene October 6 and 7th in Boston MA
- <http://biomedhpc.med.harvard.edu>



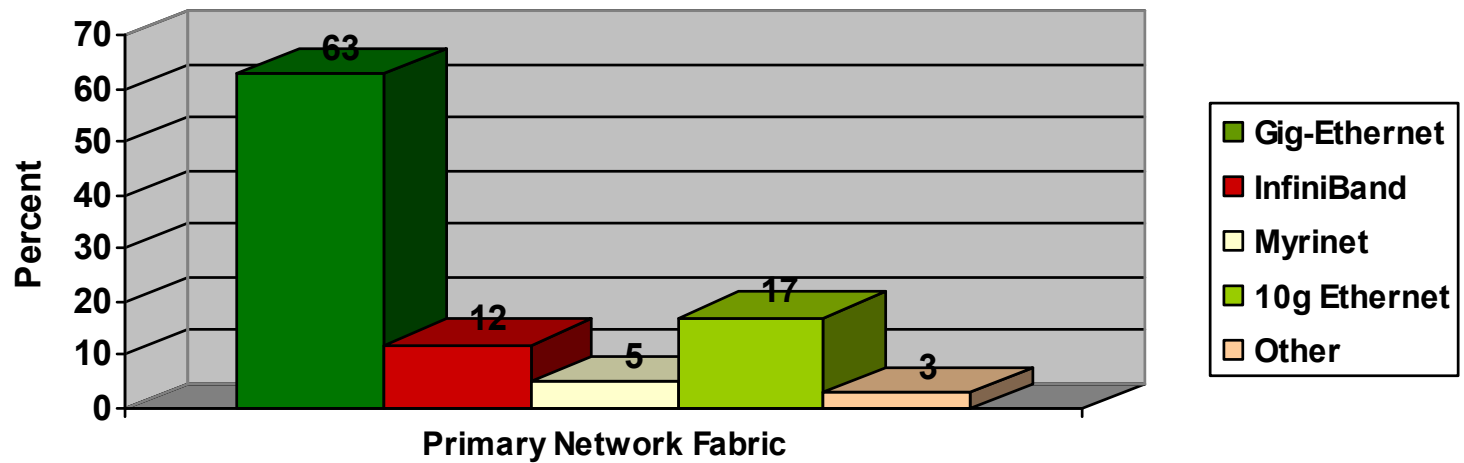
Biomed HPC Audience Surveys

- Audience response devices
- N=60-100 Leaders in HPC
- Questions asked over the two day event
- And, survey says!



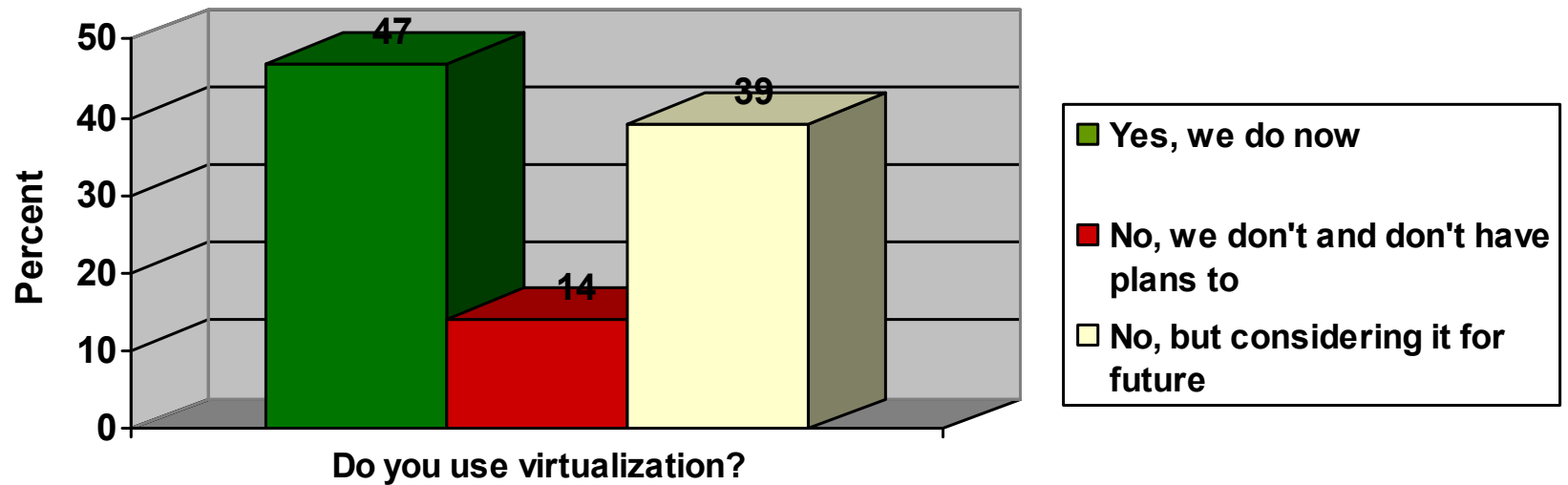
Primary Network Fabric

HMS Biomed HPC Leadership Summit 2007



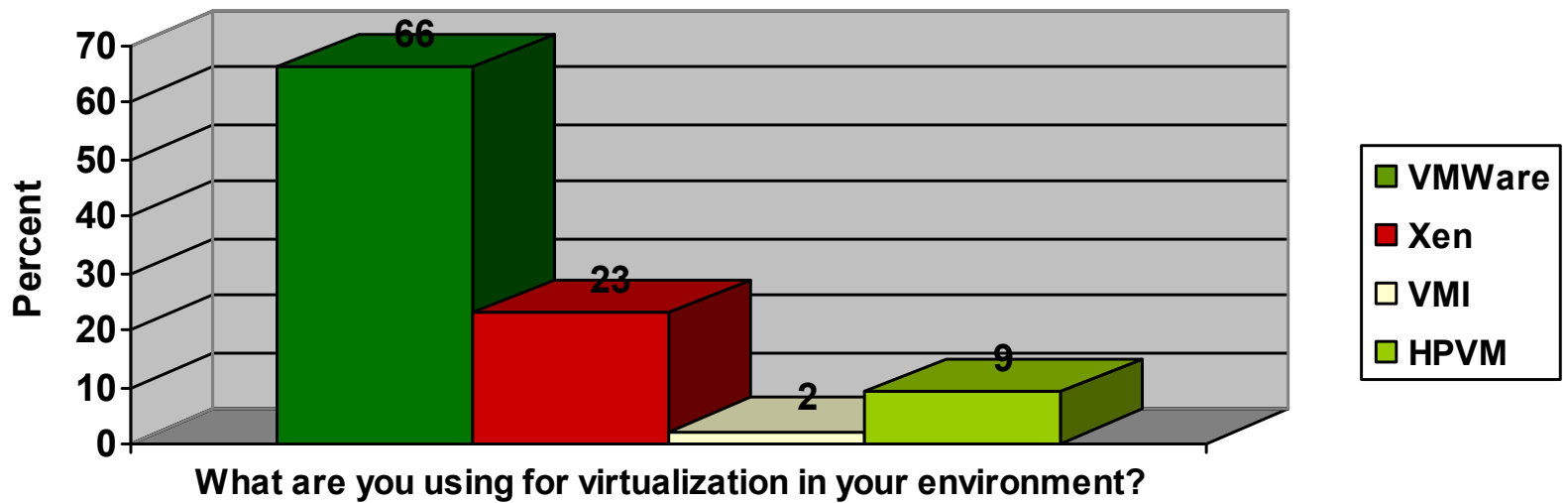
Do you use virtualization?

HMS Biomed HPC Leadership Summit 2007



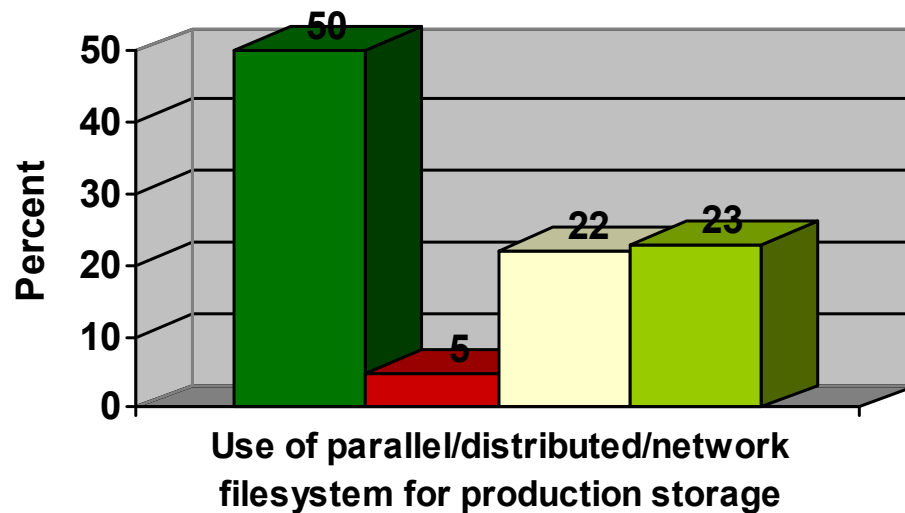
What are you using for virtualization?

HMS Biomed HPC Leadership Summit 2007



Use of parallel/distributed FS

HMS Biomed HPC Leadership Summit 2007

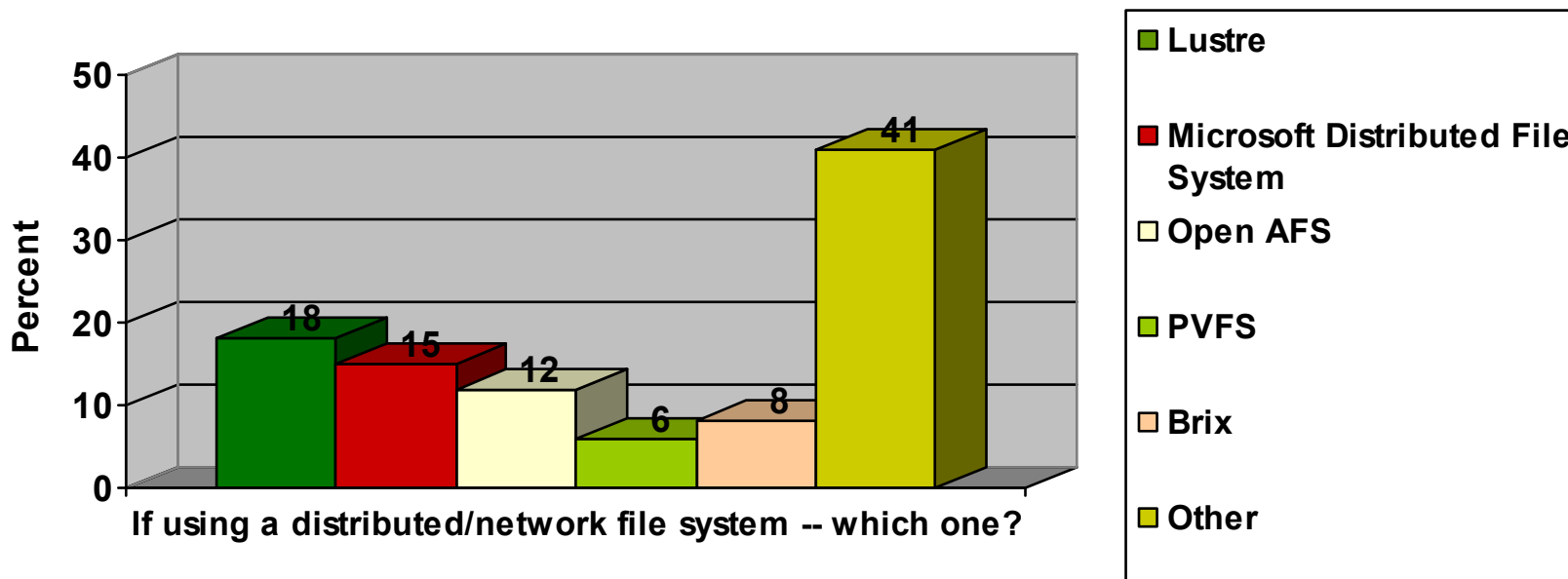


- Yes, we do now
- No, we don't and don't have plans to
- No, but have plans to
- No, but considering for future



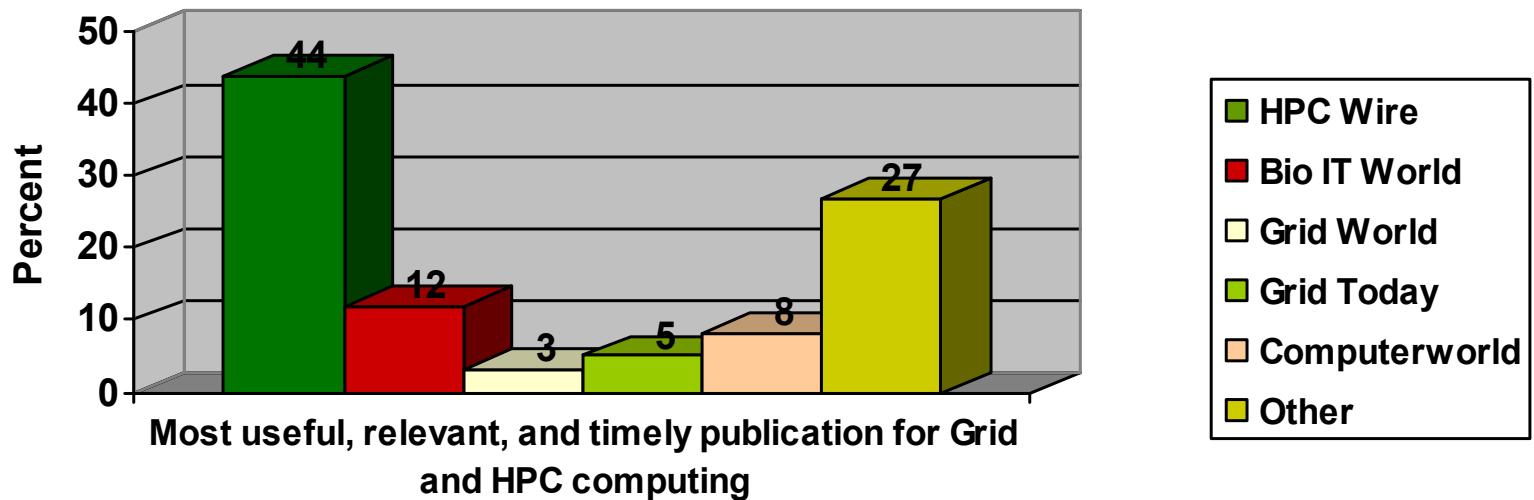
Which parallel filesystem?

HMS Biomed HPC Leadership Summit 2007



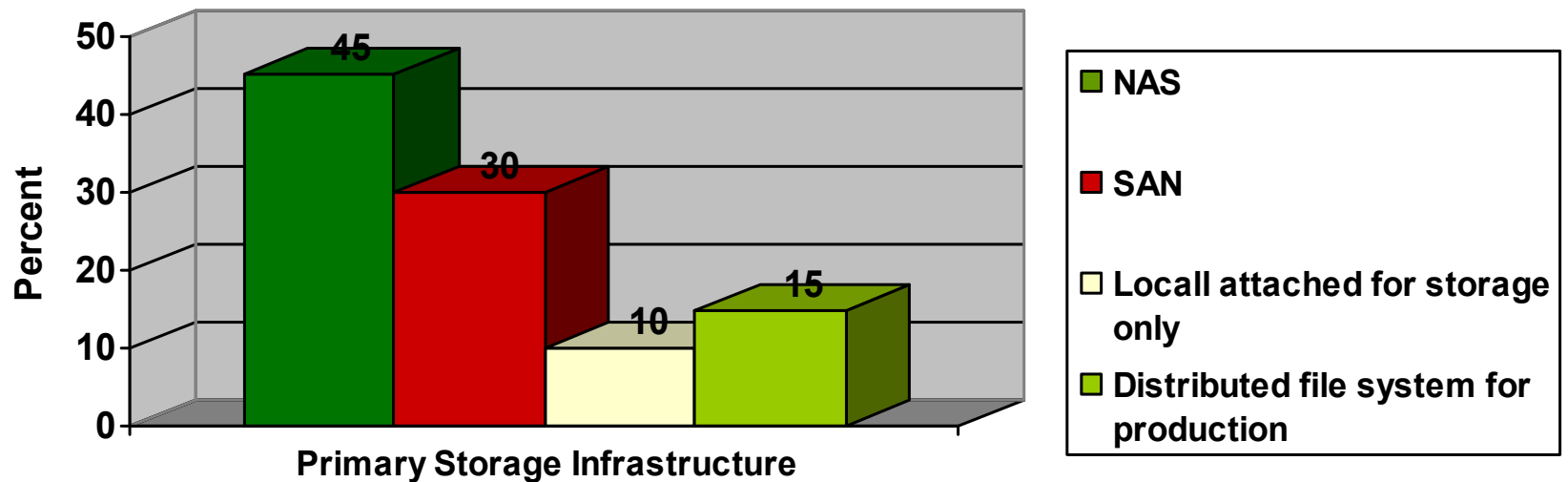
Which publication do you rely on?

HMS Biomed HPC Leadership Summit 2007



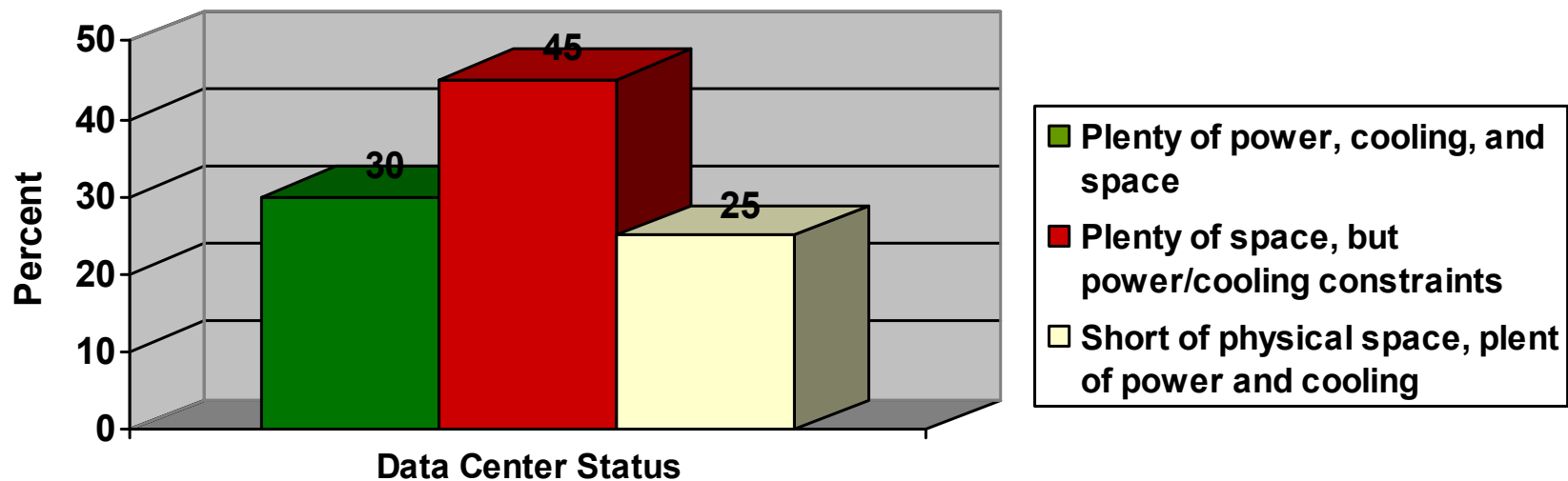
Primary Storage Infrastructure

HMS Biomed HPC Leadership Summit 2007



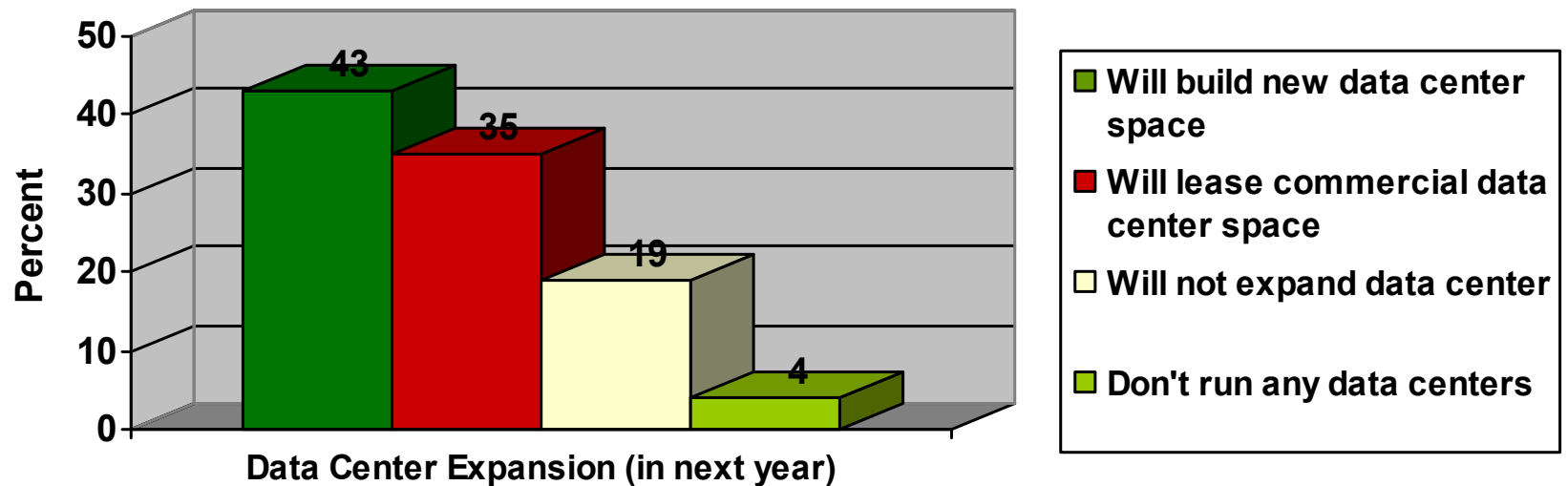
Data center challenges

HMS Biomed HPC Leadership Summit 2007



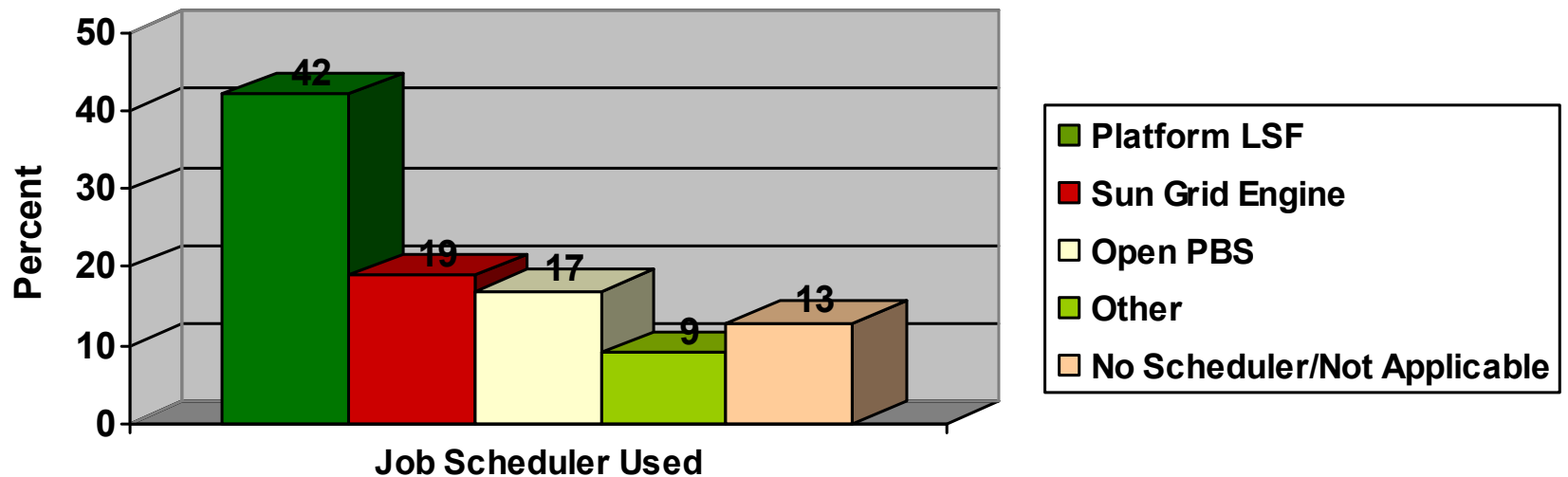
Data center expansion plans

HMS Biomed HPC Leadership Summit 2007



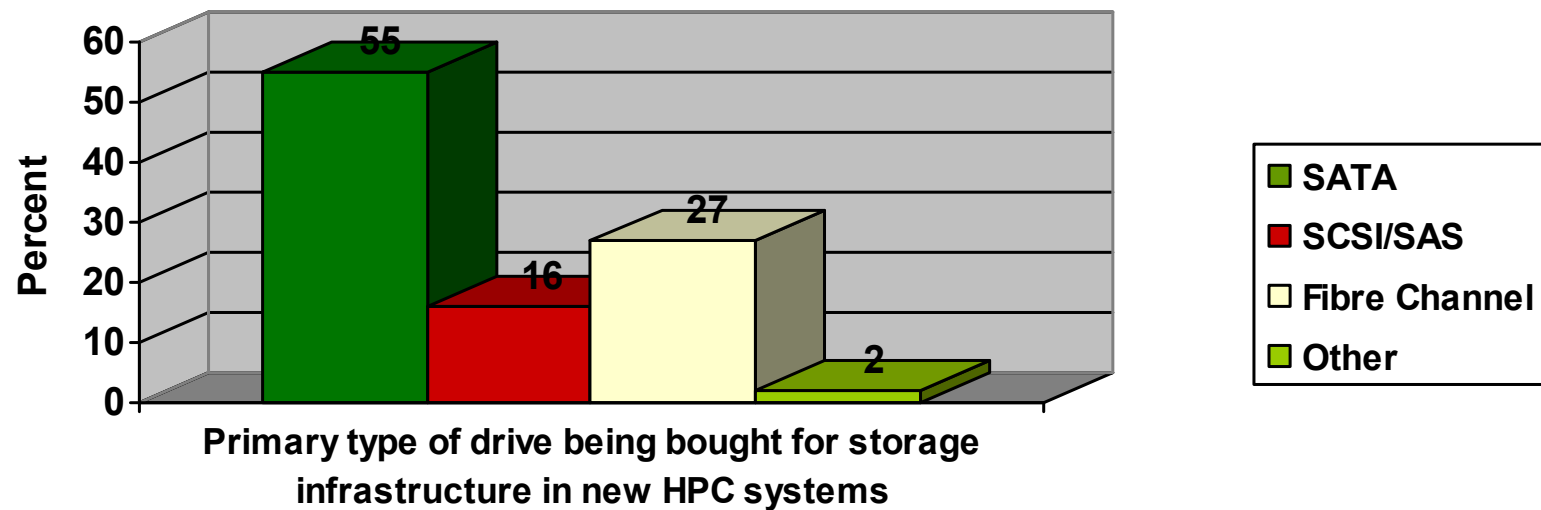
Job schedulers used

HMS Biomed HPC Leadership Summit 2007



Primary drives being purchased

HMS Biomed HPC Leadership Summit 2007



Prediction

- Biomed HPC will continue double digit growth for the foreseeable future
- The importance of open source will continue to increase dramatically
- Biomedical HPC will become more centralized



Recommendations

- User centered design
 - End to end analysis of products usability and integration
- Keep it simple
- Encourage enlightened self interest
- It's the community, stupid



Thank you

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

