Bringing Back the “Super”
EC (Super)computing

EC/CIOB/ITID/NOCD

Carol Hopkins/Luc Corbeil

November 28-29 2007

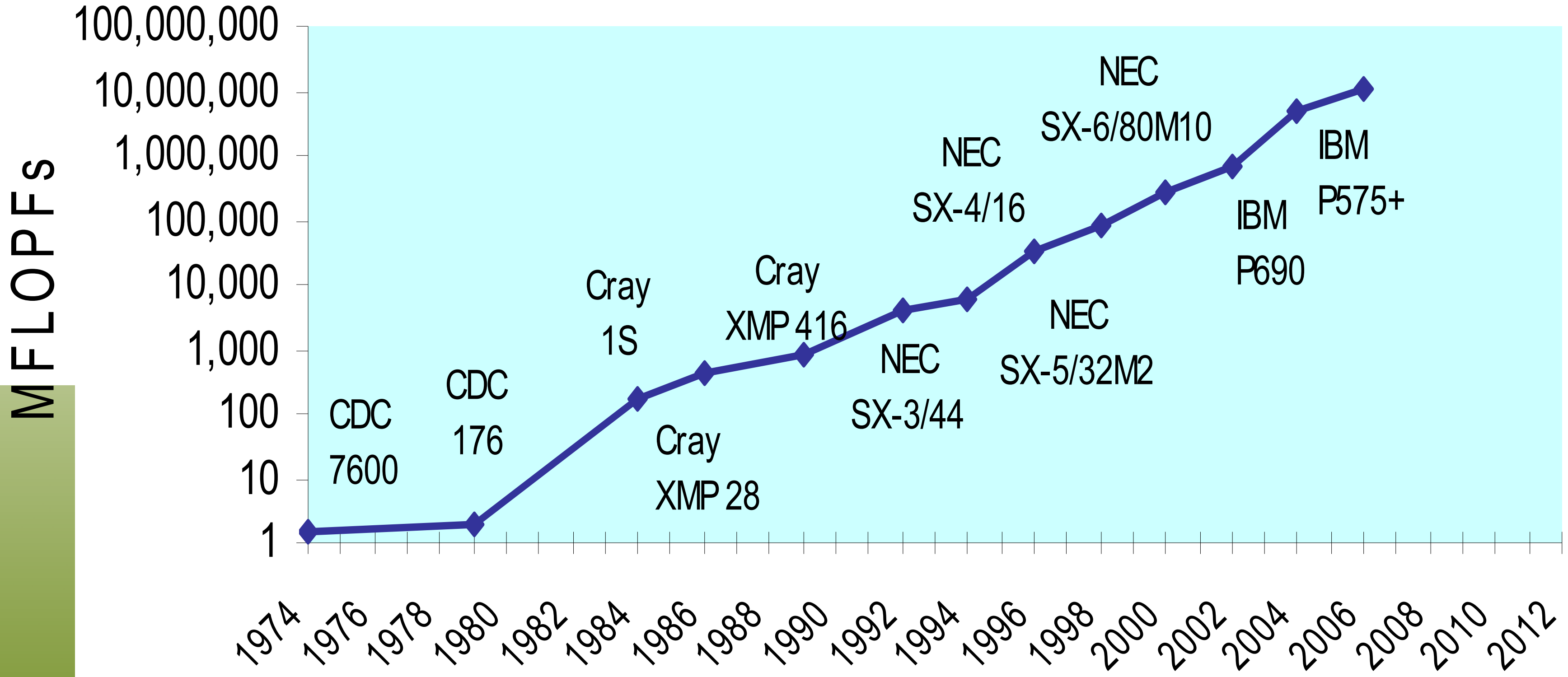
Outline

- History
- Rationale
- Upgrade
- Other key systems

History

- Supercomputer contract with IBM started in Dec. 2003
  - Azur, p690, 960 cpus, Power4 1.3GHz
- Upgrade 1 was delayed by 6 month, accepted Dec. 2006
  - Maia/Naos, p575+, 1424 cores, Power5+ dual-core 1.9GHz
- Upgrade 2 is optional
  - 30 month after acceptance of Upgrade 1, i.e. June 2009
  - Will bring us to Dec. 2011
  - If not exercised, RFP!
- RFP needed for 2012 and beyond

History


MFLOPFs
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
CDC
7600
CDC
176
Cray
1S
Cray
XMP 28
Cray
XMP 416
NEC
SX-3/44
NEC
SX-4/16
NEC
SX-5/32M2
NEC
SX-6/80M10
IBM
P690
IBM
P575+
1974
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012

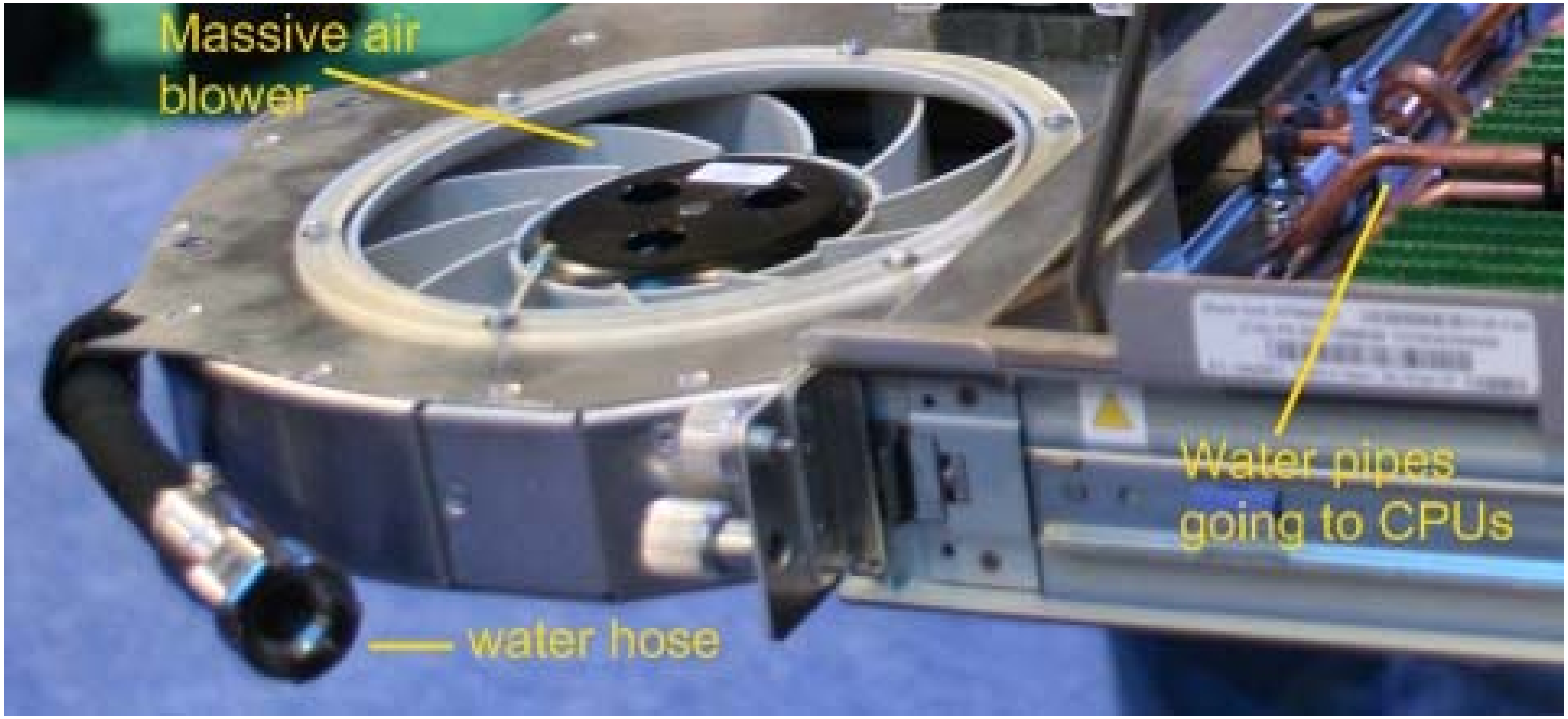
Rationale

- Discussions started with IBM last spring about exercising option for Upgrade 2.
- EC expressed needs:
  - More computing power
  - Sooner than contract timeline
  - No downtime
  - No extra funding
  - Electrical power infrastructure limited

Rationale (2)

- IBM suggested to stay with Power5+
  - Power6 quite different
    - Water cooled
    - Cpu architecture
    - Codes performance
  - Migration effort for user codes is nil
  - Installation process is known and well documented
  - More of the same for sys-admins (just more!)
  - Price of p575+ hardware going down, can give more flops
  - Meets all our contractual requirements

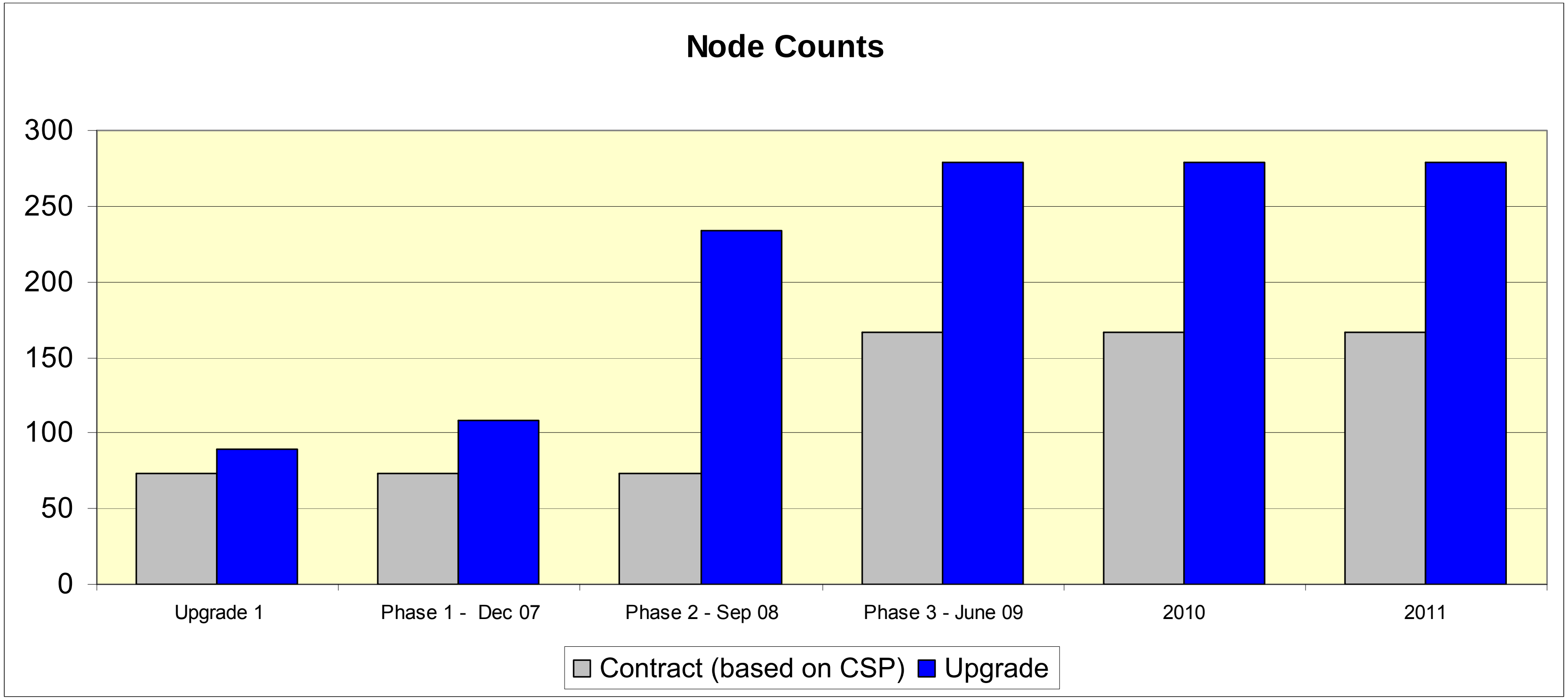
Power6: water cooled



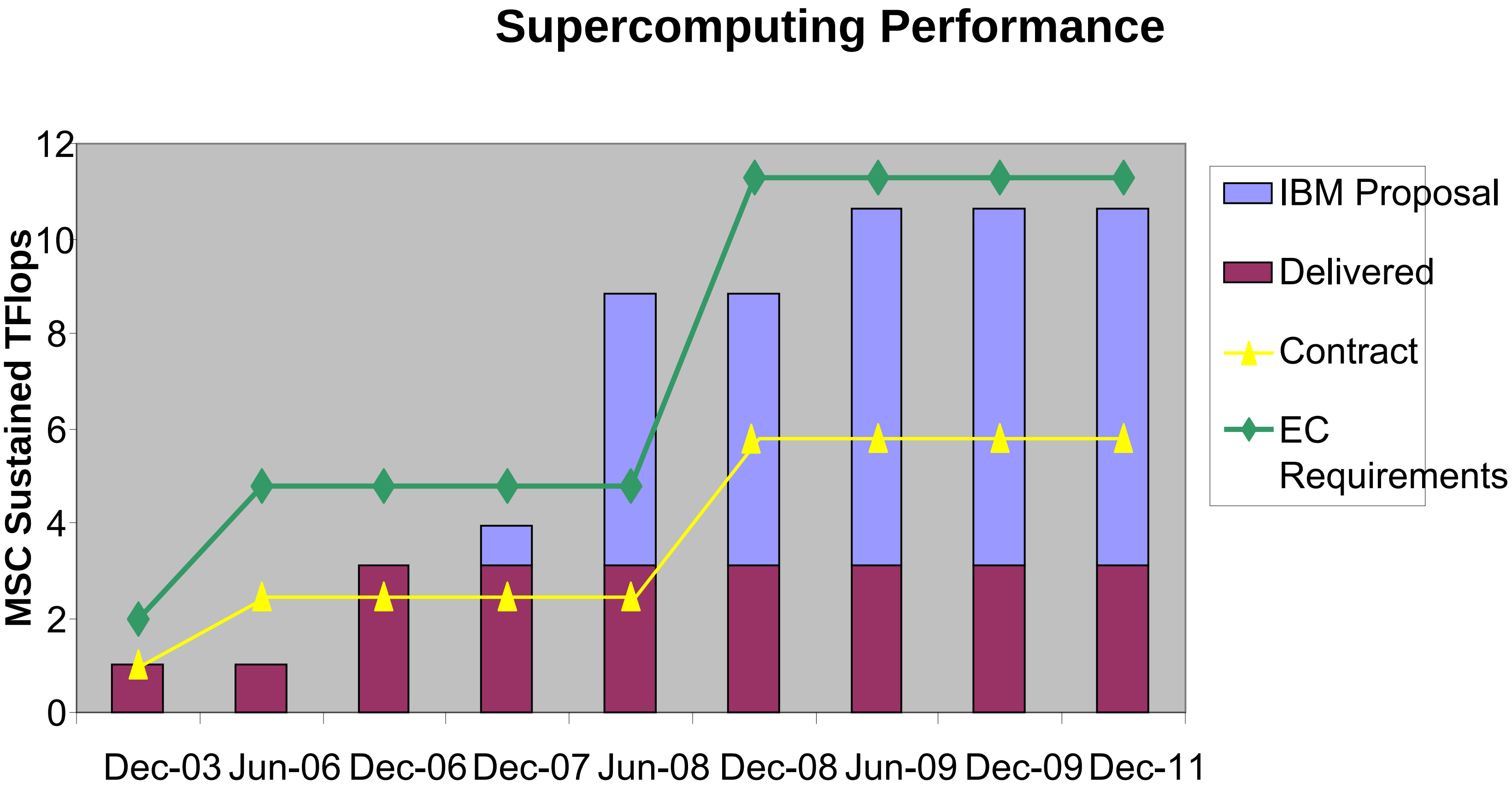
Rationale (3)

- Pros
  - Avoid heavy RFP costs (human and $$)
  - More computing power sooner can materialize
  - Upgrade path much easier for users (and a bit easier for us!)
- Cons
  - Same technology for 5 years
  - Hardware failures likely to increase
    - IBM still committed to 99% availability
    - Redundancy level is high
  - Serial or quasi-serial applications will not see a gain

IBM Upgrade


Node Counts
300
250
200
150
100
50
0
Upgrade 1
Phase 1 - Dec 07
Phase 2 - Sep 08
Phase 3 - June 09
2010
2011
Contract (based on CSP)
Upgrade

Performance Curve



Total Node Counts by Phase

Project Phase	Delivery Dates	Additional Nodes	Total node count	Total Contract nodes (based on CSP)
Upgrade 1 – Phase 1 and 2 (Maia / Naos)			91	73
Upgrade 2 - Phase 1	Dec 2007	19	110	73
Upgrade 2 - Phase 2	Sept 2008	127	237	73
Upgrade 2 - Phase 3	June 2009	45	282	167

Note: Total Node count does not include 2 nodes purchased by EC in March 2007

Installation: a puzzle

Date	C4 (maia)	C5 (naos)	C6 (TBD) 128 ports	C7 (TBD) 256 ports
Actual	40	38		
Dec 2007	59	38		
Sept 2008	59	38	121	
Dec 2008	59	→	80	123
Jan 2009			80	182

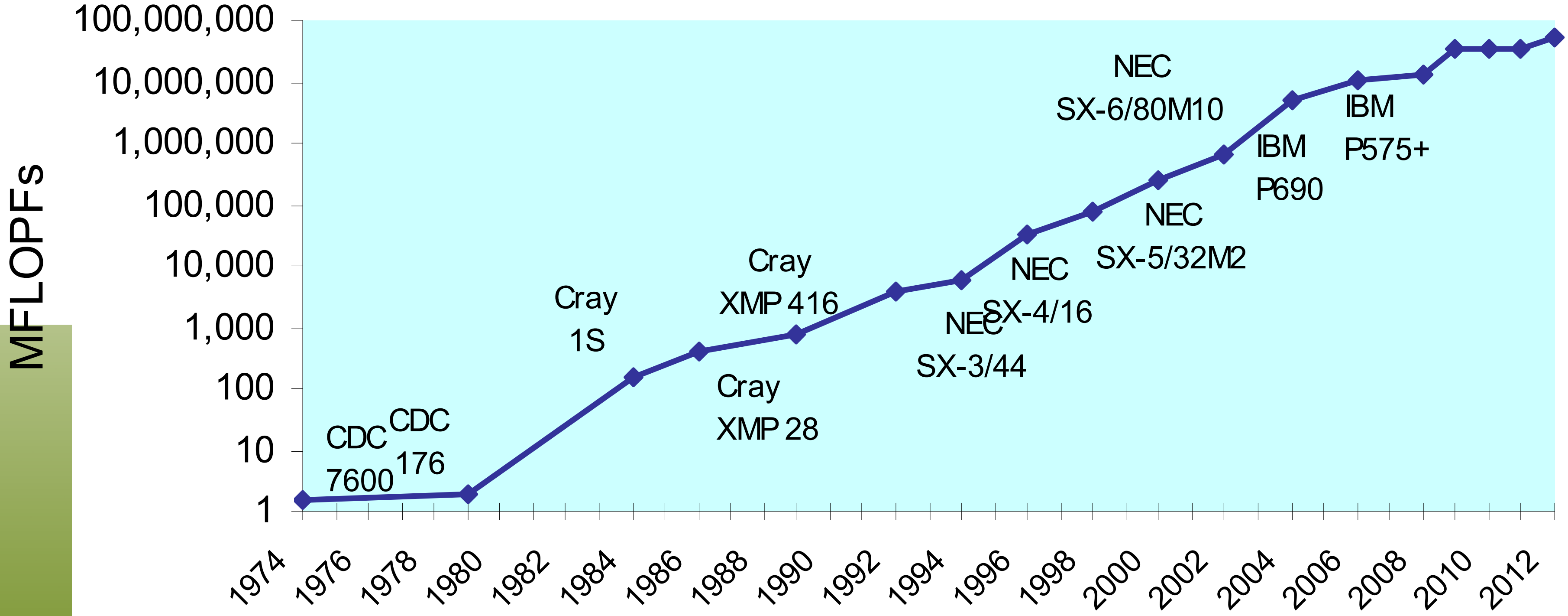
Storage, HPS, Networking

- Storage: same technology, doubled
  - From 36 to 72 TB raw, ~ 54 TB configured
  - Easily meets requirement on paper
- HPS: tripled
  - From 2*64 to 128 + 256 ports.
- Networking: doubled
  - From 2*6513 to 4*6513 Cisco enclosures
  - Extra 10 GigE ports to connect to other equipment

Power, Cooling

- The project is power-driven!
  - 19 extra nodes: waiting for power, due mid-December
  - 121 nodes cluster: waiting for power, matter of weeks
  - Not everything will be on UPS (compute nodes on dev side)
- Cooling
  - Avoided plumbing that Power6 required
  - Additional Lieberts will be installed
    - Not only for the supercomputer

Outlook



Environnement Canada Environment Canada

Canada

IBM Upgrade

- 67% more than contractual requirement. Why?
  - Marketing: to keep EC as a customer
  - Marketing (2): they hope to get 1st place on Top500 in Canada
    - With extra 19 nodes: #400 (Nov. 2007)
    - University of Sherbrooke: #391
  - To avoid cost and effort of a RFP
  - To ensure guaranteed revenue until 2011
  - To occasionally access a p575+ test system next spring

IBM other opportunities

- Could expand clusters
  - About 100 extra compute nodes can be “easily” added
- Discussion to test newer technology
  - Power6 blades (not water cooled)
  - Cell blades (CPU similar to a PlayStation)
- Closer partnership with other IBM sites

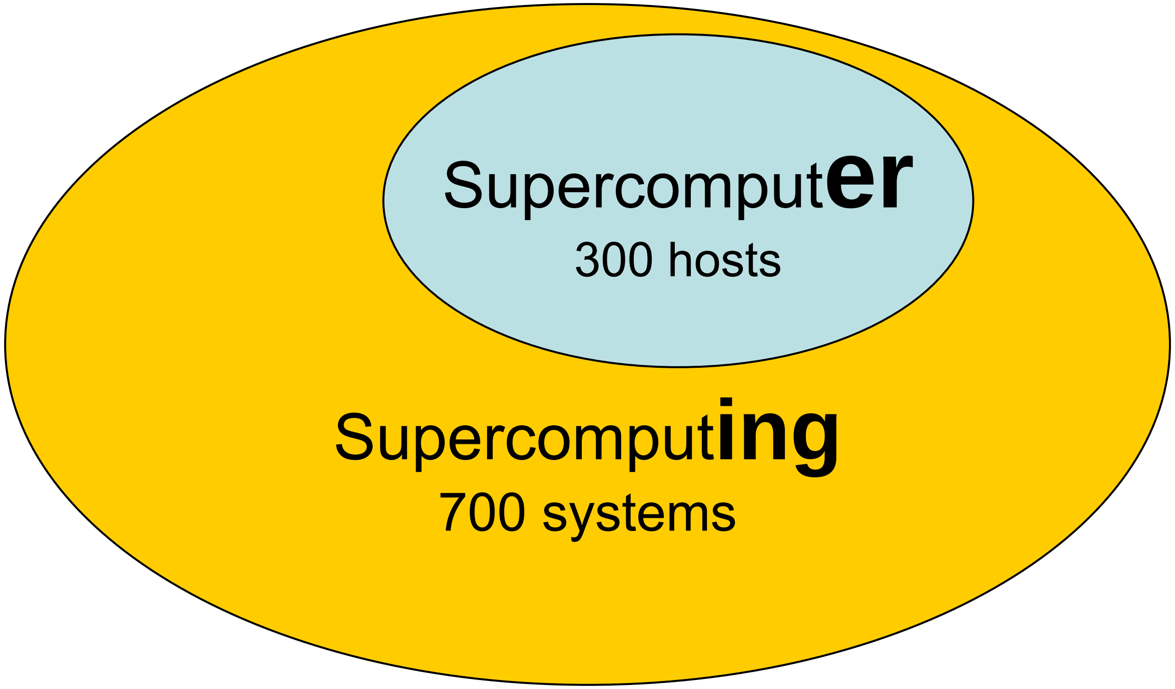
What we will try to provide

- Smooth transition
  - No file migration, disk space should “magically” expand
  - No significant downtime
  - Newer OS (AIX 6.X) and compilers (xlf 11.X)
- Give extra computing resources as soon as possible
- Better monitoring of resource usage/training
  - Require more staff!

What is expected from users

- Continuous feedback
  - Both bad and good
  - Can’t live without it!
  - *DDS for problems*
  - *it_infrastructure_client_feedback@ec.gc.ca* for feedback
- Patience
  - Operations first!
  - No power, no computer…
- Willingness to test new OS/Compiler
  - Ideally, could be close to automated
- Think about the future
  - MPI jobs of O(10k) or O(100k) processors as soon as 2012!

IT Infrastructure



Supercomput**er**

300 hosts

Supercomput**ing**

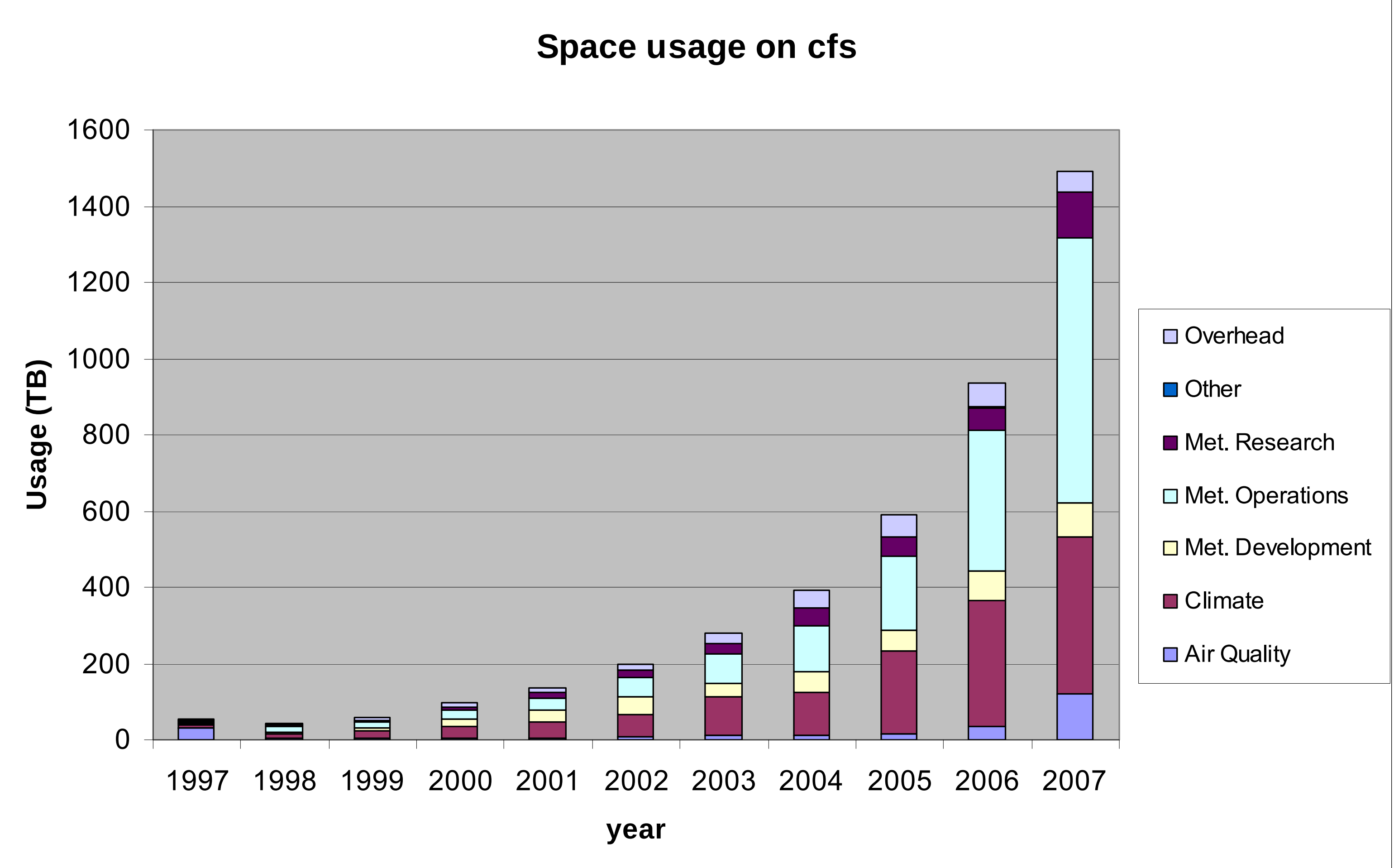
700 systems

Front-ends

- SGI O3900 (castor and pollux)
  - Pollux: 40 processors 600 MHz, 40 GB RAM
  - Castor: 32 processors 1 GHz, 32 GB RAM
  - Over 40 TB of high-performance disks
    - TP9500, 2GBit
    - IS4500 (equiv. TP9700), 4GBit

CFS: Archiving

CFS Growth


Space usage on cfs
Usage (TB)
1600
1400
1200
1000
800
600
400
200
0
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
year
Overhead
Other
Met. Research
Met. Operations
Met. Development
Climate
Air Quality

New HSM software

- Was ADIC (now Quantum) Fileserv/Volserv
  - Support ended by Quantum Sept. 2007
- Now Quantum StorNext
  - Based on Fileserv
  - LTO-3 tapes and drives (2X faster/bigger)
- Migration
  - New SGI Altix 350 CFS server
  - Must migrate files from Fileserv/oldcfs to StorNext/cfs
  - Migration 80% completed

Ib-dorval

- 40 nodes, dual Xeon processors
- Split in two, 20 nodes for ops
- Operational!
- Will replace gfx (finally)
- Infiniband switch, 4X, 900 MB/s
- Storage: Rackable RapidScale (4TB currently, 40 TB when power and manpower permits)

Next steps

- Front-ends and CFS
  - Contract ending Feb-March 2009 (one opt. year)
  - Replacement project starting January 2008
  - Installed by end 2009
- New UPS spring 2010
  - Interesting puzzle!
- Supercomputer RFP starting 2009, acceptance fall 2011
- IBM contract ends Christmas 2011

Thank you!

- Questions?
- Feedback is welcome: luc.corbeil@ec.gc.ca