BaBar Databases

Jacek Becla

Stanford Linear Accelerator Center

BaBar

- Studying difference between matter and antimatter (CP violation)
- Based at SLAC
- ~600 physicists, 10 countries
- In production
  - Run 1: Oct'99 – Oct'00, ~23.5 fb^{-1}
  - Run 2: Feb'01 – Jul'02, up to ~60 fb^{-1} now

World-wide Collaboration

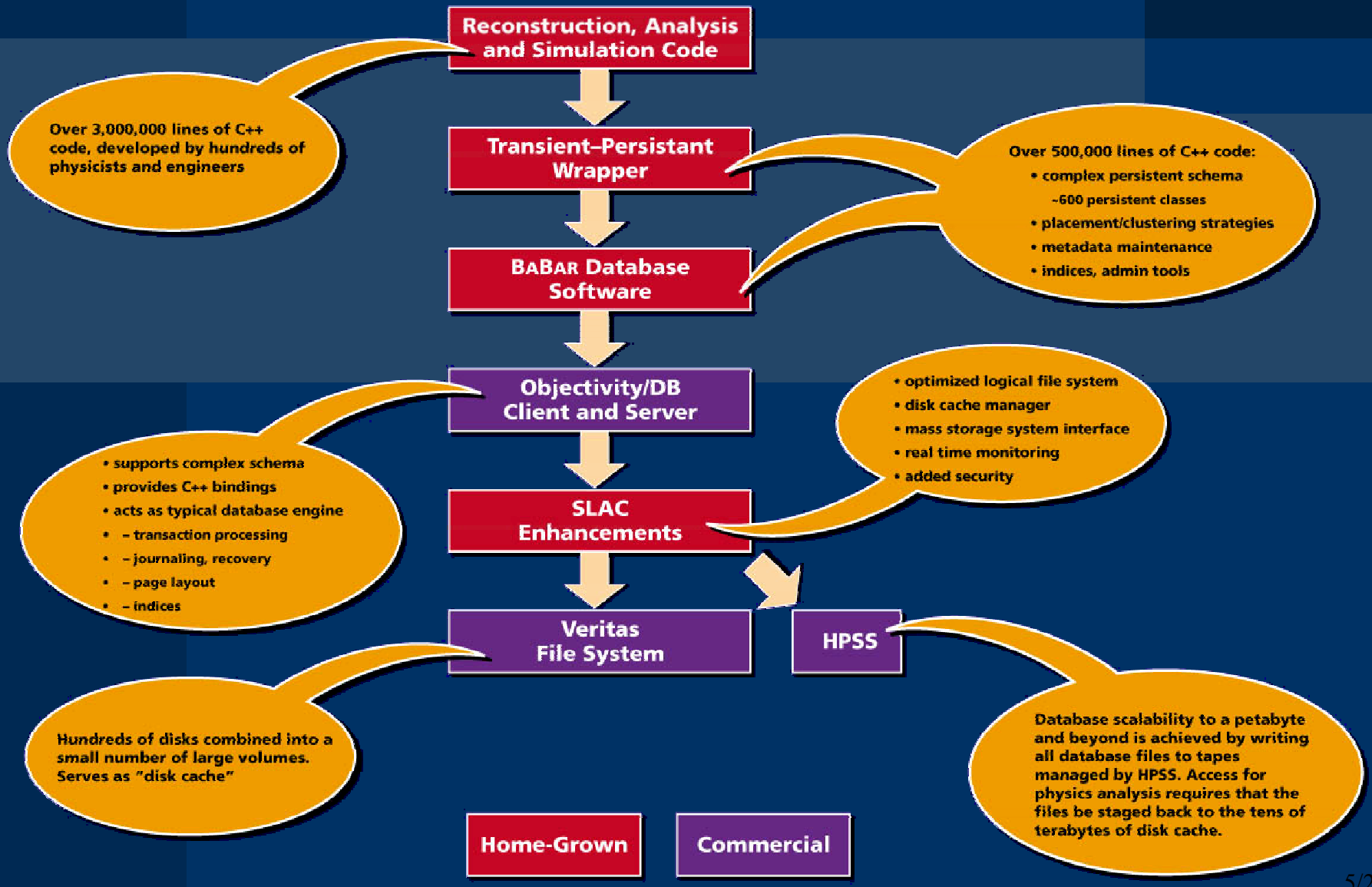
- Canada, China, France, Germany, Italy, Norway, Russia, Taiwan, UK, USA
- Tier A centers
  - SLAC, INFN, IN2P3, RAL
  - IN2P3 already in operation
    - 2001: 130 CPUs, 25 TB, 155 Mbits/s
    - 2002: 380 CPUs, 45 TB, 622 Mbits/s
- Tier C centers
  - Caltech, Colorado, many others
  - Number of Tier C centers expected to rapidly grow

BaBar Tier-A Centers

Annual Growth

	2001	2002	2003	2004	2005
CCIN2P3 Tier-A Disk (TB)	15	30	20	20	30
CCIN2P3 Tier-A CPU (30 Spec95 units)	100	280	400	800	1000
INFN Tier-A Disk (TB)		13	16	23	27
INFN Tier-A CPU (30 Spec95units)		390	480	690	810
UK Tier-A Disk (TB)		20	20	30	40
UK Tier-A CPU (30 Spec95 units)		200	300	450	500

BaBar Software



Production

- Per subsystem
  - Prompt Reconstruction @ SLAC
    - 8 4-CPU Sun servers, ~200-300 clients
  - Reprocessing
    - Moving to Padova, starting July ‘02
  - Analysis @ SLAC
    - ~30 2 and 4-CPU servers, ~200 clients
  - Simulation
    - outside SLAC, 15 sites now
- HPSS (SLAC and IN2P3)
  - Automatic staging used at IN2P3
  - on demain staging used at SLAC
- Veritas FS

Hardware

- Some numbers
  - Total # servers @SLAC
    - 80 data servers
    - 43 lock servers
    - 25 jnl and catalog servers
  - Disk space for objy data: ~60 TB
  - All servers: Sun, Padova first to try Linux servers
- Total hardware cost ~$30M
  - tapes and manpower not included
  - Driving cost: disks, then CPU

Bdb Responsibilities

Provide efficient way of managing BaBar's data

- Development
  - Customizing underlying database engine to fit BaBar's needs
    - Based on ODBMS: Objectivity/DB
    - ~500K lines of code on top
- Administration
  - Running system in production
  - User support
  - Data distribution

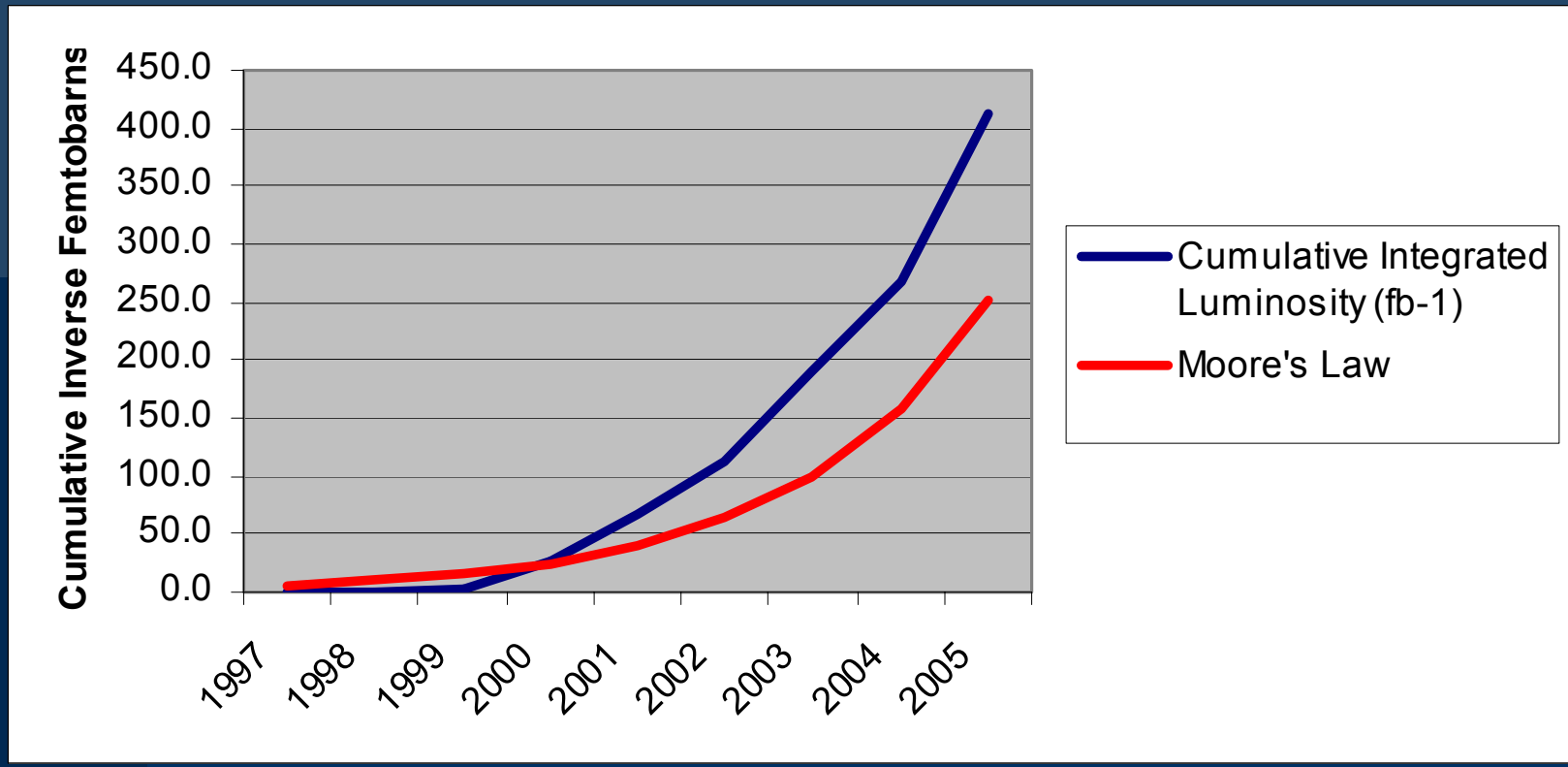
History

- Production started in Oct'99
  - Initial design requirements (100 Hz, 200 TB/year) met 1 year after the start
  - Many unexpected issues, bottlenecks
  - Changing physics requirements
  - Challenge to keep up with improving accelerator performance
- Improving, optimizing and tuning, redesigning, scaling
  - since day one, still ongoing
    - E.g. reducing event size now

Interesting Challenges

- Scaling the system with always increasing luminosity
- Optimizing the system
  - Packing data, getting most out of the hardware
- Improving robustness
  - More data -> more hardware -> more failures
- Indexing frequently accessed data
- Plugging into Data GRID

BaBar Computing Outlook



Data Distribution - History

- 130+TB of SP4 data imported 2 SLAC
- 80+ TB exported to IN2P3
  - Regularly 0.3 TB/day
  - Export from IN2P3 to SLAC
- Both Tier A and C export data
- Distributed effort
  - Germany, UK, France, Italy, USA, Canada

Data Distribution - tools

- SLAC -> IN2P3: BdbServer & Jimport
- SP sited -> SLAC: MocaEspresso
- All: BdbDistTools and bbftp or bbcp
- All tools written in C++, Perl
- Automated import procedure at SLAC to cope with increased growth of SP production sites

Known Issues

- Unreadable databases
  - Due to outstanding transactions on extracted dbs
  - Few problems due to bad disks
  - Most (~98%) picked up by integrity checking at SLAC
  - Improving our QA tools now
    - This includes dedicated (faster/better) QA tool from objy release 8 (~1Q' 03)
  - Fighting data dist problems ~0.3 FTE

Collection Export: BdbServer

- Used to ship objy data from IN2P3 and SLAC to external sites
- Used to ship individual collections
  - Pointer or deep-copy collections
- In general ok, very few problems

Future: Data production

- SP (till 12'03)
  - Same data volume (not shipping raw or sim, but adding more sites. ~ 1.5-2TB/week to SLAC)
- REP (Padova, Italy)
  - Dedicated server at SLAC to send/receive data
  - Mods to existing automatic import tools at SLAC needed
  - New, moca-based tool developed (in Italy)
  - Initially plan to ship all data to SLAC
  - Later IN2P3 might take data directly from Padova

New Development (1)

- Redesigning BdbDistTools
  - Ease of maintainance
    - Almost non-maintainable now
  - Ability to plug in grid applications
  - Ease of use at Tier A, C sites
  - Expected to be completed ~3Q'02

New Development (2)

- BdbServer++
  - Goal: to allow users to create and extract deep copies of pointer collections
  - To be done
    - Integration of BdbCopyJob with BdbServer
    - Bridge fd phase III
    - RDBMS for collection-db mapping
  - Timescale: 3-4Q'02

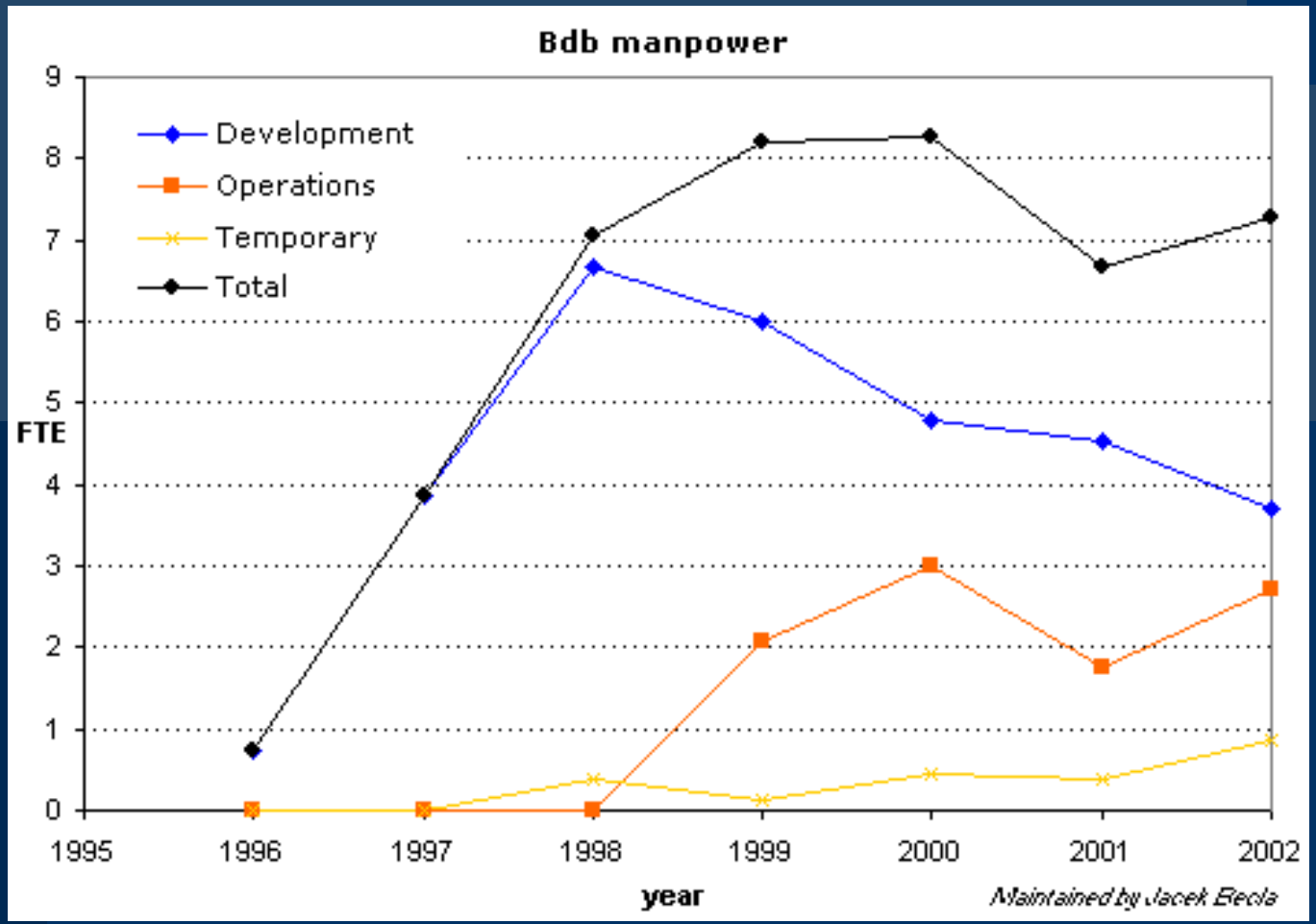
New Development (3)

- Bridge fd phase III
  - Input fd != output fd
  - Allows to extract a collection from one fd, and store it in a different federation
  - Scheduled July 15, 02
  - All done, in tests now, already successfully shipped some test data to IN2P3
- QA tools

The Largest

news and articles in many places, including CNN, Computerworld, Information Week, KRON 4, The Australian

Manpower



Total = ~45 FTE-years

ODBMS…

- Objectivity
  - Excellent support, new features, improvements
- ODBMS not growing as fast as expected
- LHC likely to use Hybrid Solution (ROOT+RDBMS)

Summary

- Per system
  - OPR, REPRO scales well now
  - Analysis, serious effort to optimize (speed and event size) started recently
  - SP done at remote sites. A success
- Large focus on data distribution issues now
  - Distributed effort
  - Should try to converge to one global solution
- More at http://www.slac.stanford.edu/bdb