D0 Computing Retrospective

Amber Boehnlein

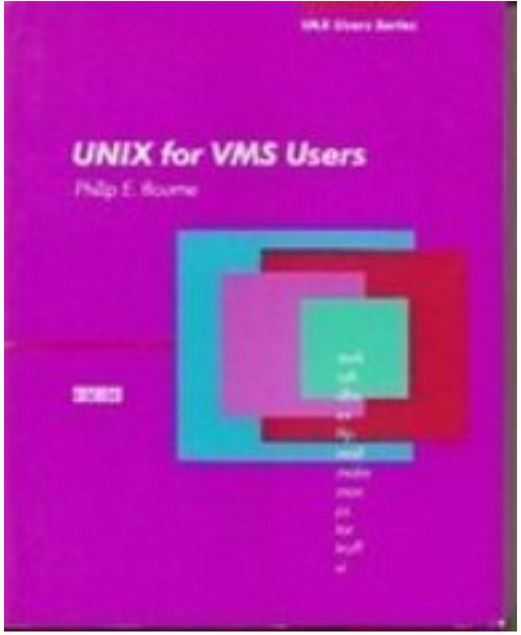
SLAC

June 10, 2014

This talk represents 30 years of outstanding technical accomplishments from contributions from more than 100 individuals.

Run I Computing

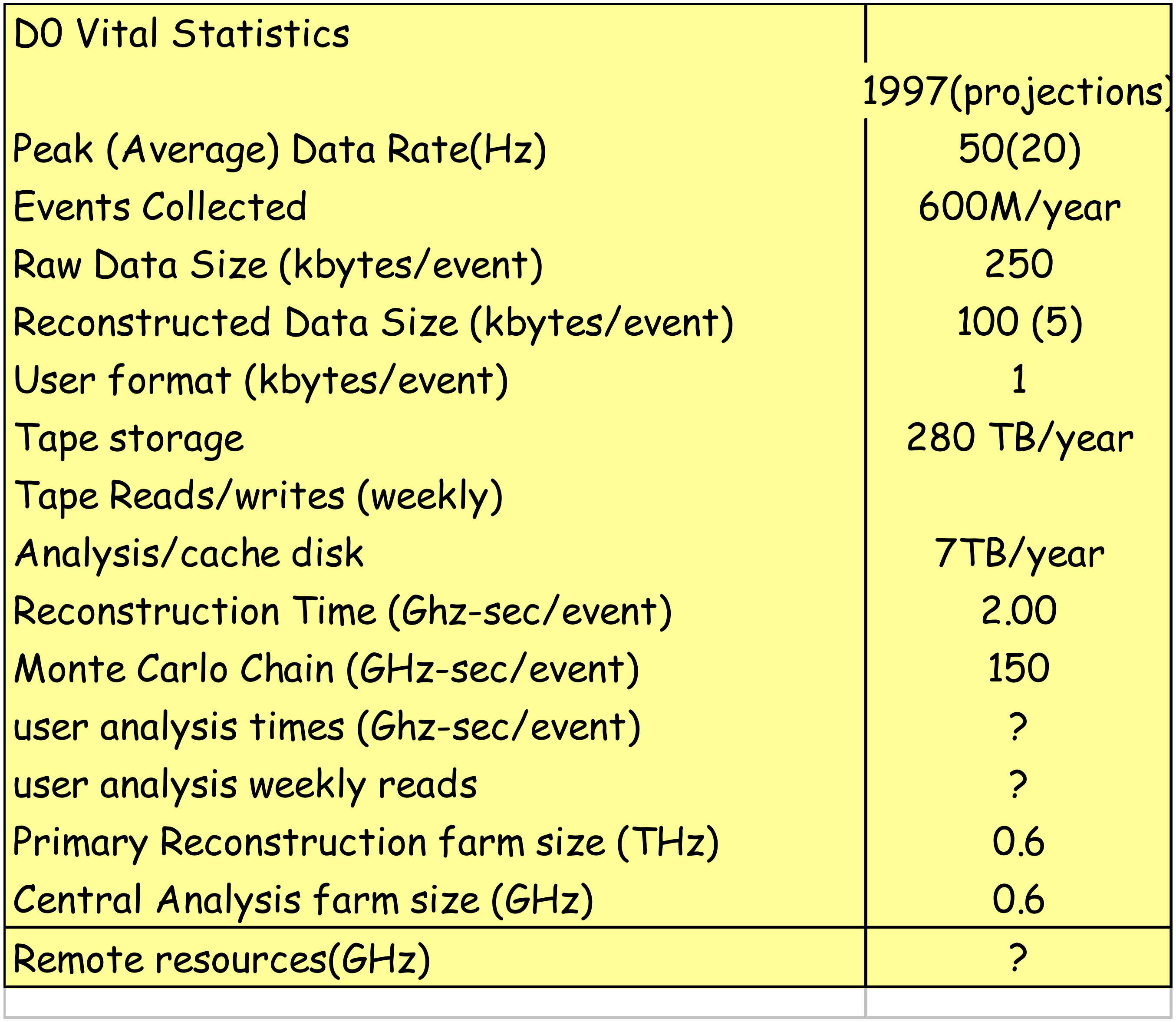
- VAX, VMS and Fortran ruled the day
  - Some computing in the porta-camps would trip off
  - Transition to UNIX…
- Limited resources == compromises
  - Baby sitting jobs
- Fatmen was a rudimentary data management system
- Command line interfaces
- Mike Diesburg and Qizhong Li were the go-to folks!



Run II Planning: 1997

- Planning for Run II computing was formalized in 1997 with a reviewed bottoms-up needs estimate.
  - Critical look at Run I production and analysis use cases
- The planning started with vision of what a modern computing and analysis system should do and how users should interact with the data.
- The planning for the LHC MONARC Model and BaBar Computing was roughly concurrent
  - There was no C++ standard
  - Computing architectures were in transition
- Tight budgets for hardware and software projects
  - The FNAL CD, CDF and D0 launched on a set of Joint Projects.

Statistics 1997



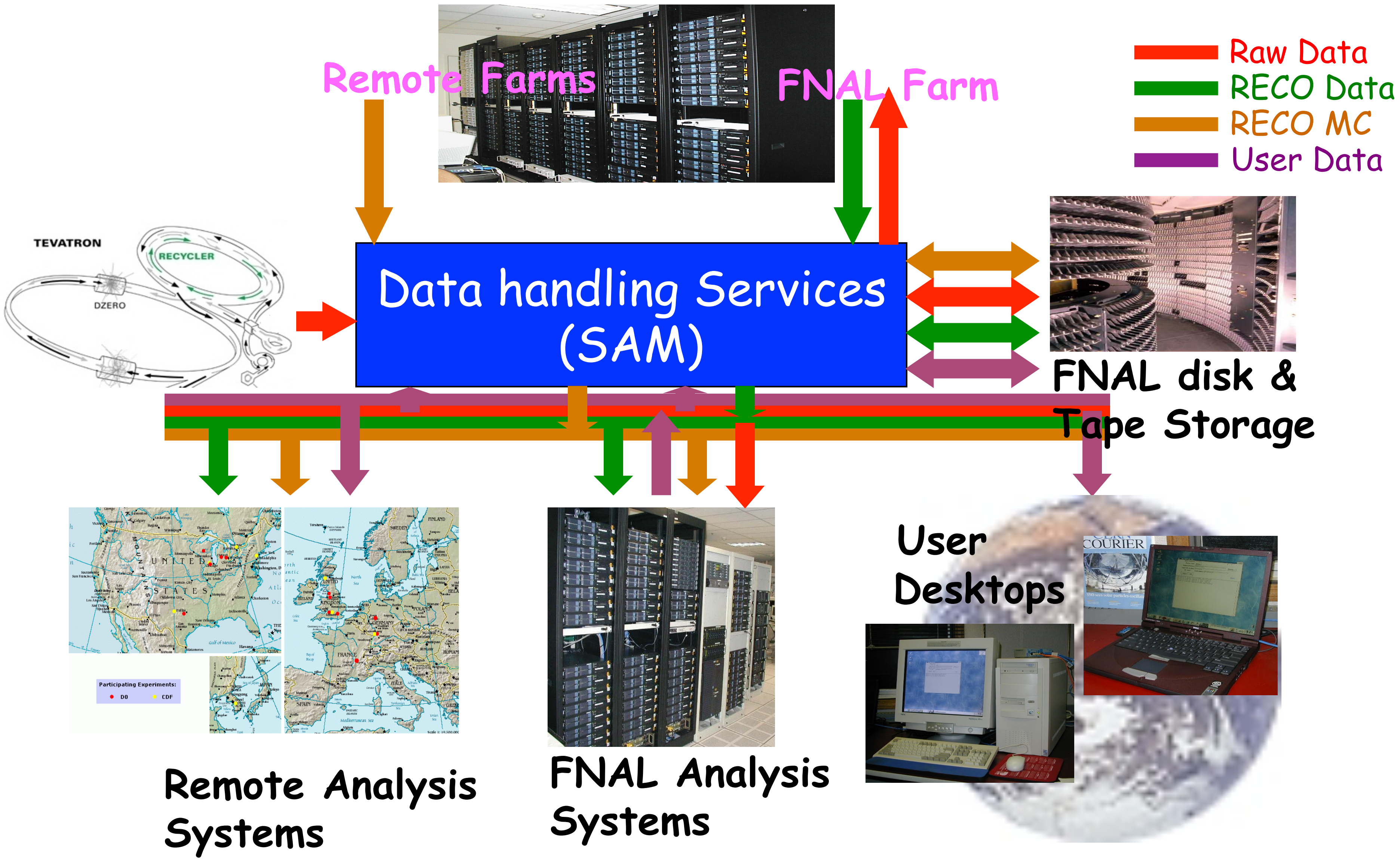
D0 Vital Statistics	
	1997(projections)
Peak (Average) Data Rate(Hz)	50(20)
Events Collected	600M/year
Raw Data Size (kbytes/event)	250
Reconstructed Data Size (kbytes/event)	100 (5)
User format (kbytes/event)	1
Tape storage	280 TB/year
Tape Reads/writes (weekly)	
Analysis/cache disk	7TB/year
Reconstruction Time (Ghz-sec/event)	2.00
Monte Carlo Chain (GHz-sec/event)	150
user analysis times (Ghz-sec/event)	?
user analysis weekly reads	?
Primary Reconstruction farm size (THz)	0.6
Central Analysis farm size (GHz)	0.6
Remote resources(GHz)	?

In "then year" costs, much computing was a formidable challenge!

Commodity systems not in general use.

Decided to Generate MC data offsite

1997 Computing Model



🟥 Raw Data
🟩 RECO Data
🟧 RECO MC
🟪 User Data

Remote Farms

FNAL Farm

Data handling Services (SAM)

FNAL disk & Tape Storage

Remote Analysis Systems

FNAL Analysis Systems

User Desktops

SAM Data Handling

- Data volumes implied a model with intelligent file delivery to use cpu, disk and tape resources effectively.
  - Implies caching and buffering
  - Implies decision-making engine
  - Implies extensive bookkeeping about usage in a central database
  - Implies some centralization
- Consistent interface to the data for anticipated global analysis
  - Transport mechanisms and data stores transparent to the users
  - Implies replication and location services
- The centralization, in turn, required client-server model for scalability and uptime and affordability.
  - Client-server model then applied to serving calibration data to remote sites…
- Anticipated concepts: Security, Authentication and Authorization
- In production since 2001

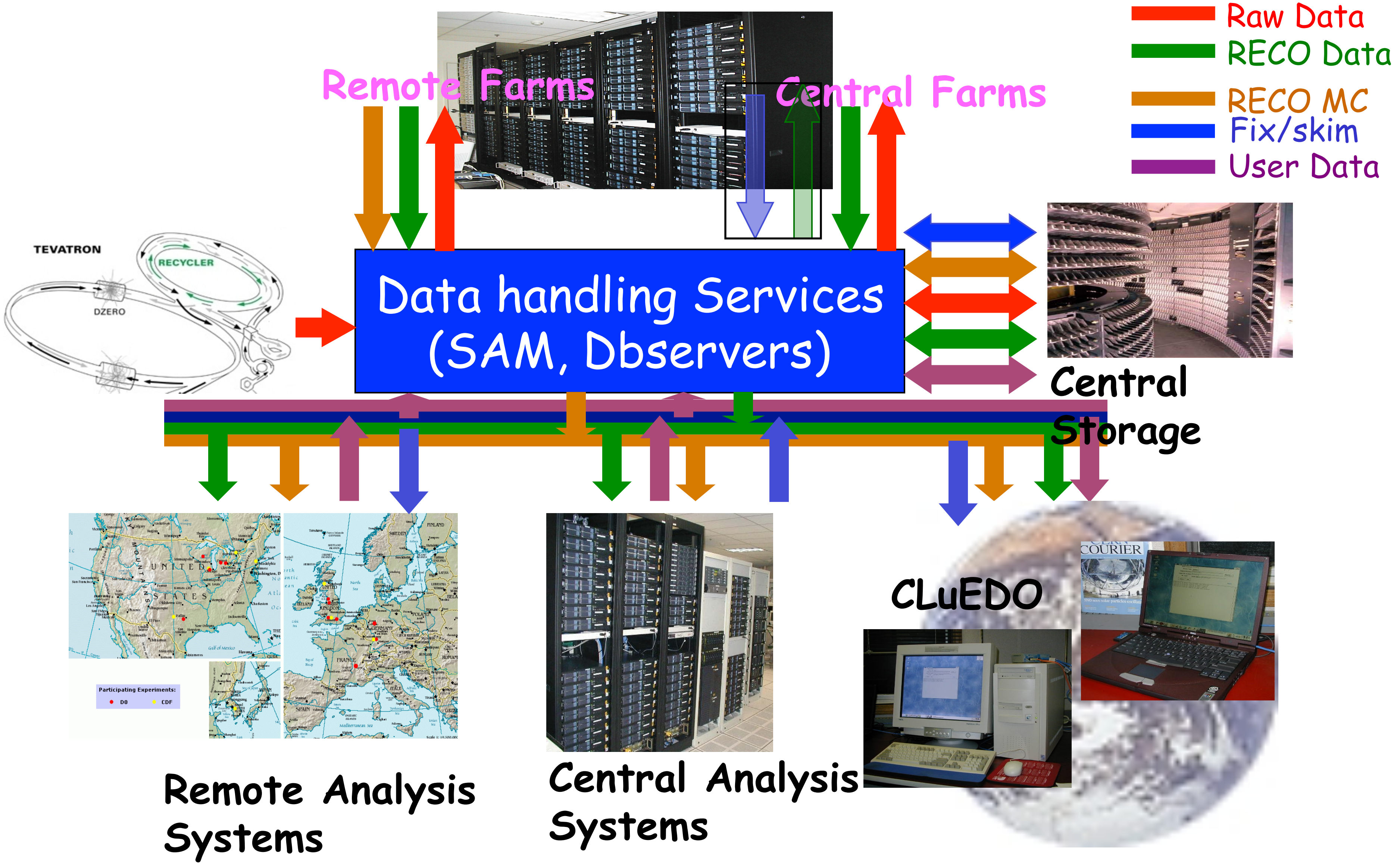
CLUED0

- 1999: A Cluster of 1 became a Cluster of 2
- Fairshare batch system on a clustered desktops managed by young physicists
  - This can only be crazy, unless it's brilliant
  - It became the backbone of the analysis computing
- Many firsts in D0 computing happened on CLUED0
  - Local builds were much faster than on SGI
  - Deployed PBS
  - First Linux SAM station was on ClueD0
  - Paved the way for the Central Analysis Backend (CAB)

Start Up: 2001-2002

- The D0 detector rolled in March 2001
- Computing was in good shape
  - Data went to tape and more importantly came back off
  - SAM had basic functionality
  - D0mino was running
  - Clued0
  - Reco Farm was running

D0 Goes Global



🟥 Raw Data
🟩 RECO Data
🟧 RECO MC
🟦 Fix/skim
🟪 User Data

Remote Farms

Central Farms

Data handling Services (SAM, Dbservers)

Central Storage

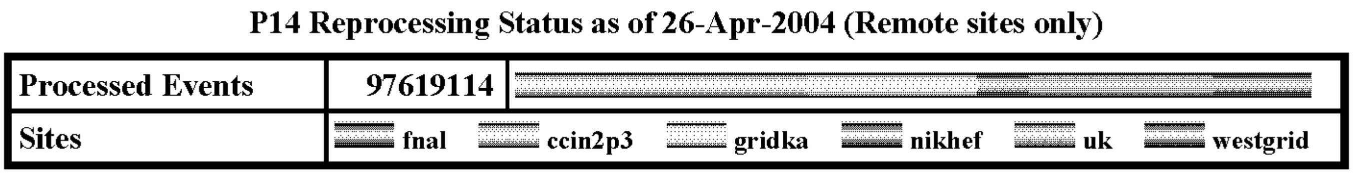
Remote Analysis Systems

Central Analysis Systems

CLuEDO

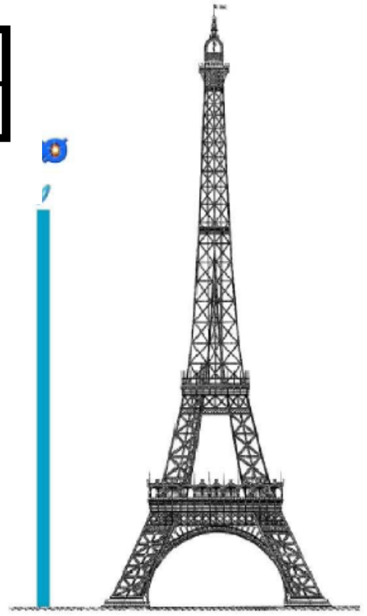
The first reprocessing

- 2003 “DST” Reprocessing with “p14”—first “global” data production: 3 months preparation: six weeks of processing
  - SAM Data Handling
  - Grid Job Submission did not working
  - 100M/500M reprocessed offsite.
  - NIKHEF tested Enabling Grid E-science (EGEE) components



P14 Reprocessing Status as of 26-Apr-2004 (Remote sites only)

Processed Events	97619114	▨
Sites	▨ fnal ▨ ccin2p3 ▨ gridka ▨ nikhef ▨ uk ▨ westgrid	

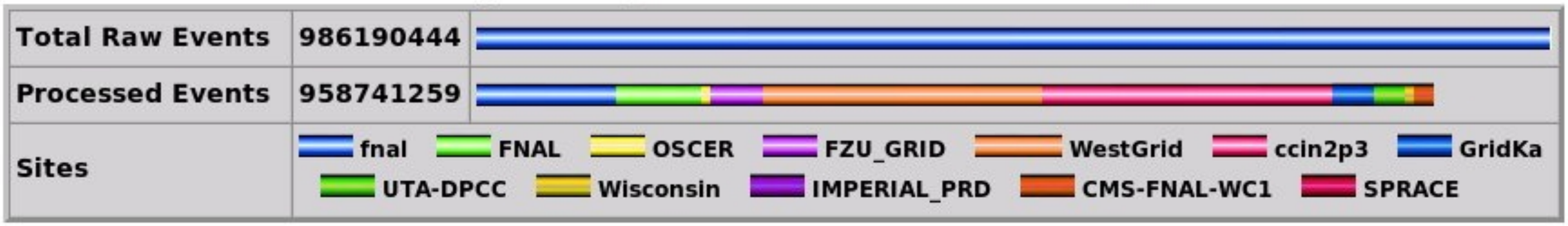


2005 Reprocessing

2005 reprocessing: Mar - Nov 05

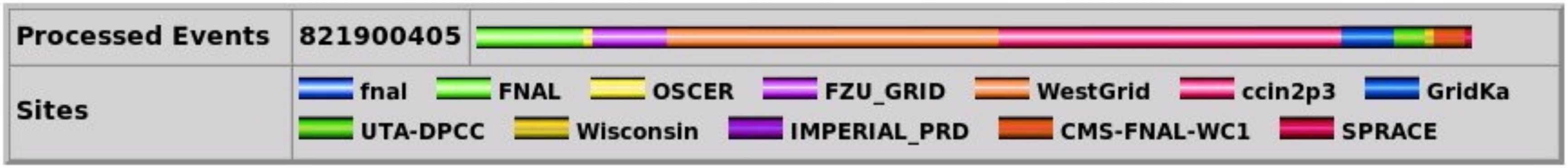
- Six months development and preparation
- 1B events from raw – SAMGrid default – basically all off-site
- Massive task – largest HEP activity on the grid
  - ~3500 1GHz equivalents for 6 months
  - 200 TB
  - Largely used shared resources – LCG (and OSG)

P17 Reprocessing Status as of 24-Nov-2005 (all sites)



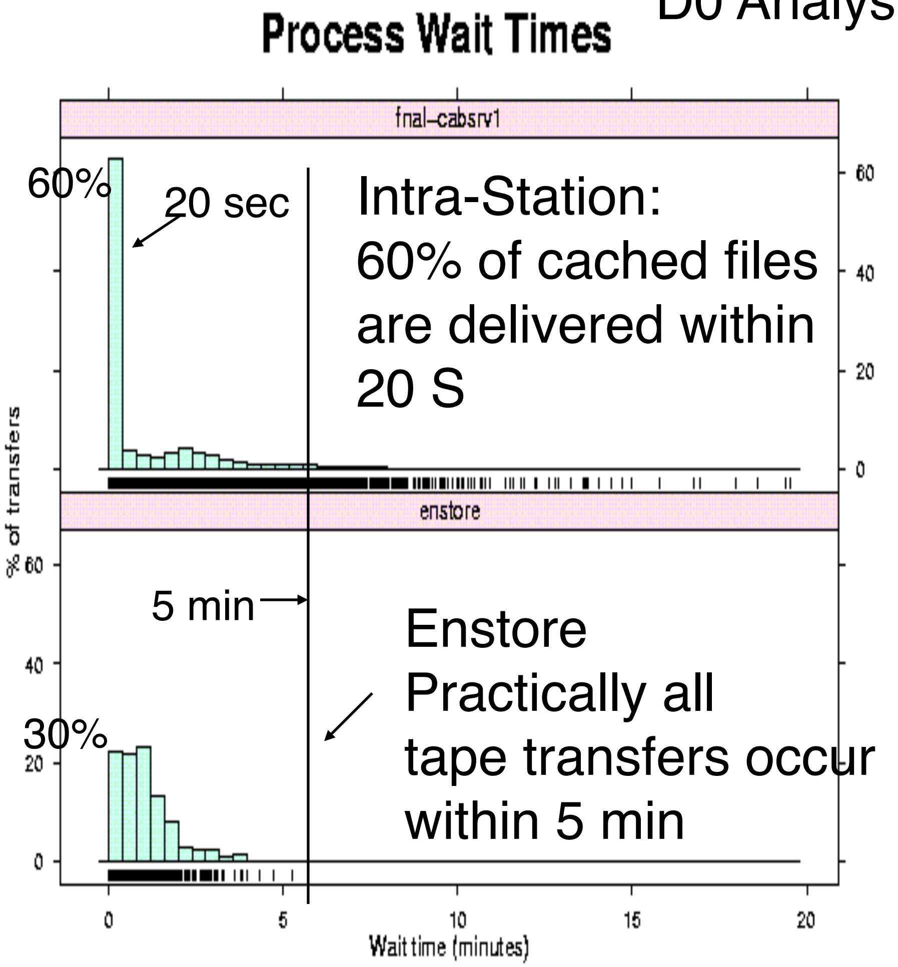
Total Raw Events	986190444
Processed Events	958741259
Sites	🟦 fnal 🟩 FNAL 🟨 OSCER 🟪 FZU_GRID 🟧 WestGrid 🟥 ccin2p3 🟦 GridKa 🟩 UTA-DPCC 🟨 Wisconsin 🟪 IMPERIAL_PRD 🟧 CMS-FNAL-WC1 🟥 SPRACE

P17 Reprocessing Status as of 24-Nov-2005 (Remote sites only)



Processed Events	821900405
Sites	🟦 fnal 🟩 FNAL 🟨 OSCER 🟪 FZU_GRID 🟧 WestGrid 🟥 ccin2p3 🟦 GridKa 🟩 UTA-DPCC 🟨 Wisconsin 🟪 IMPERIAL_PRD 🟧 CMS-FNAL-WC1 🟥 SPRACE

DO Analysis-2003



Intra-Station: 60% of cached files are delivered within 20 S

Enstore Practically all tape transfers occur within 5 min

D0 Analysis systems

User interface including batch submission –D0tools

CLUED0-managed by the users for the users

Clustered desktops with batch system and SAM station, local project disk

Developed expertise and knowledge base

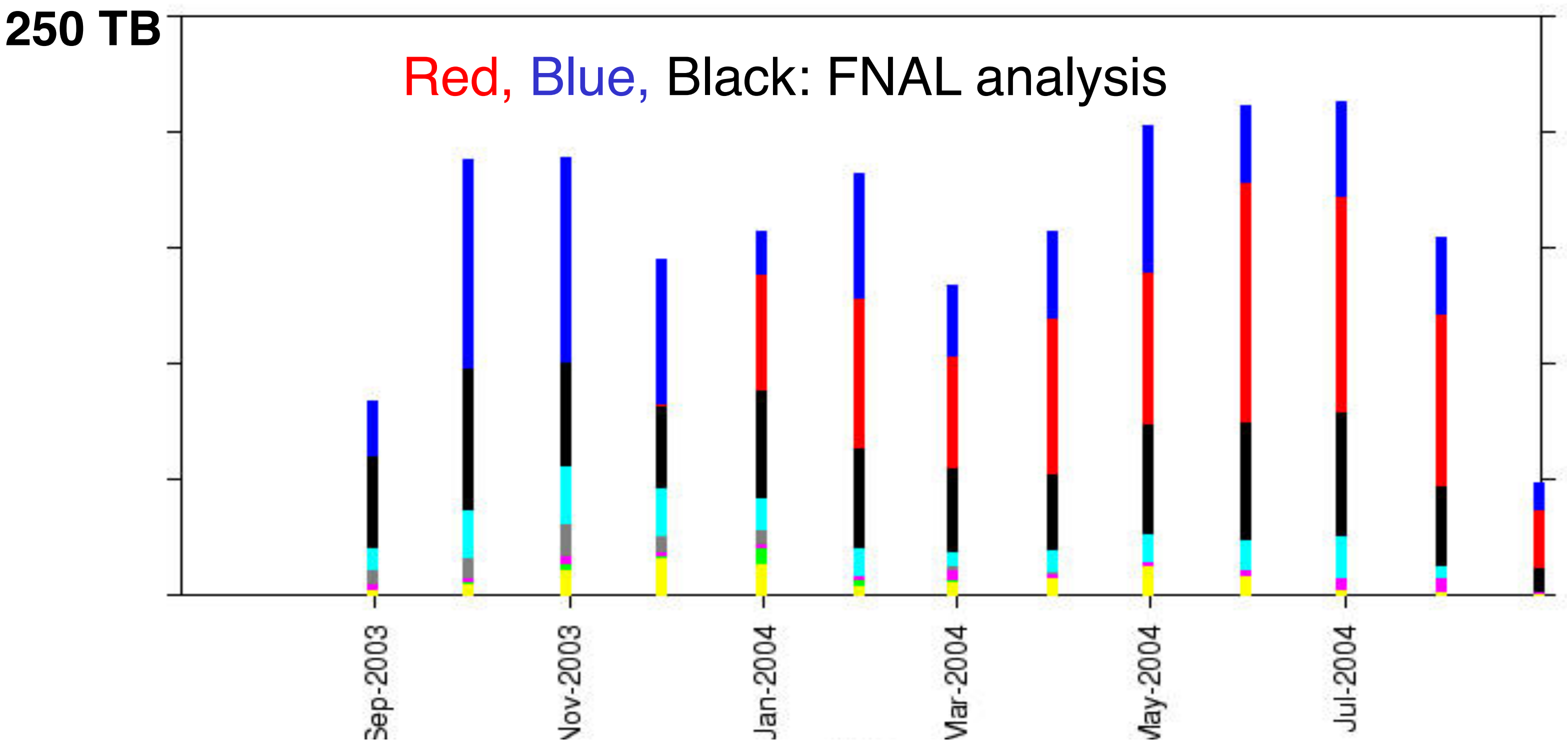
Linux fileservers and worker nodes for analysis

pioneered by CDF with FNAL/CD

Before adding 100 TB of Cache,2/3 of transfers could be from tape. Things go wrong—but also go right!

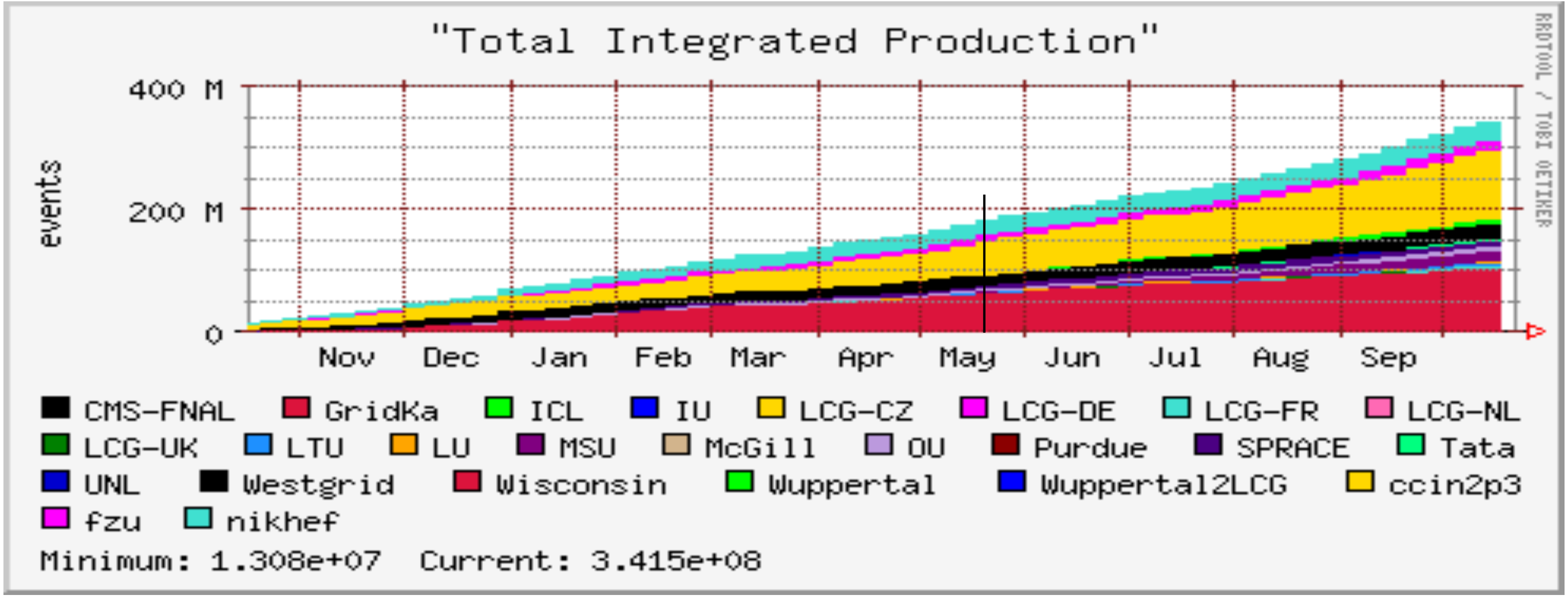
Analysis:2004

- SAM Data Grid enables "Non-FNAL" analysis
  - User data access at FNAL was a bottleneck
  - SGI Origin 2000-176 300 MHz processors and 30 TB fibre channel disk was inadequate
  - Users at non-FNAL sites provided their own job submission
  - Linux Fileservers added at FNAL—remote analysis hiatus



Monte Carlo Production

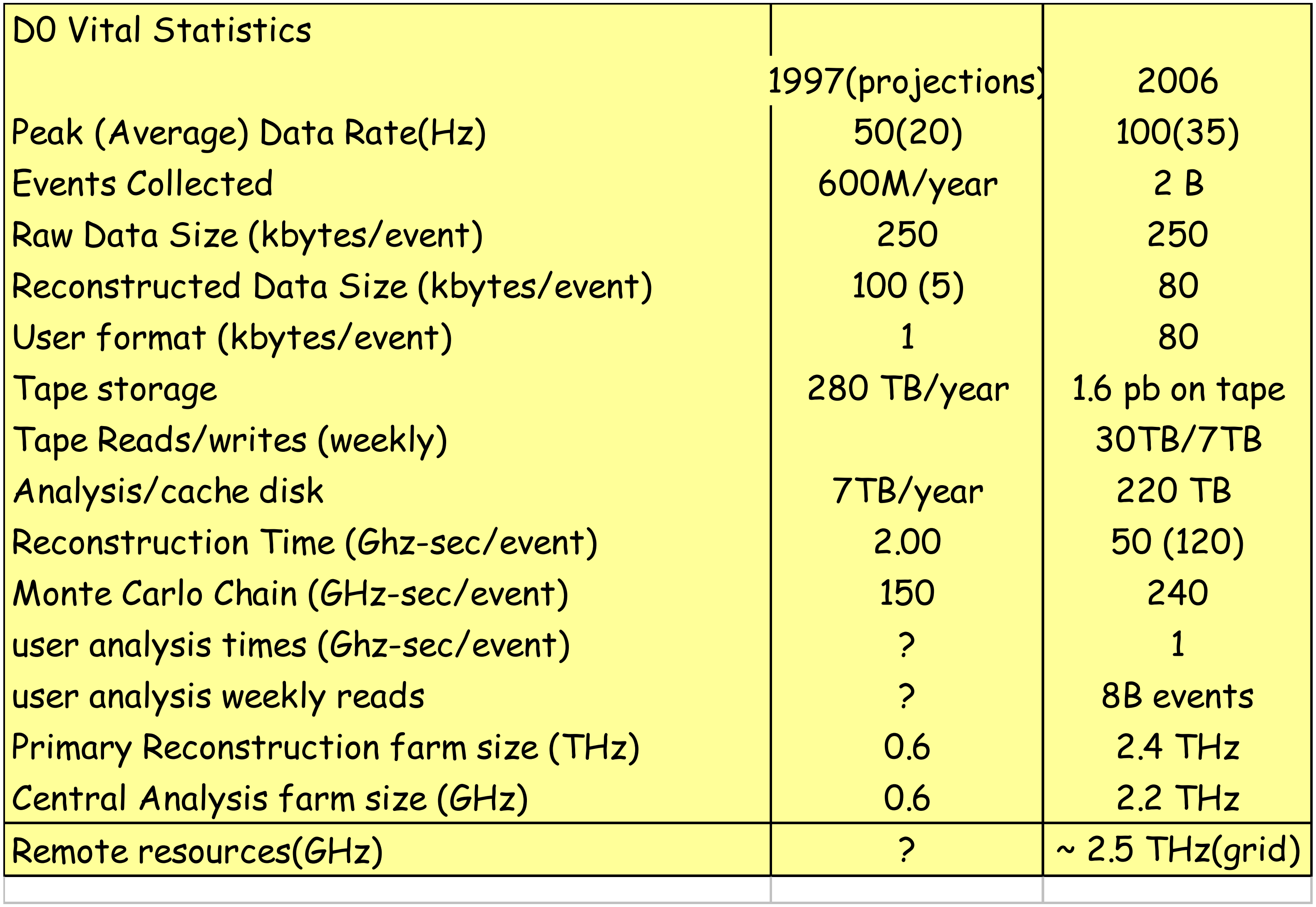
- 2004: 1M events/week peak at 6 sites
- 2006: Average 6M/week Best week 12.3 M events
- Running in “native” SAMGrid mode and in LCG interoperability mode
- Running DO MC at 6/11 LHC Tier 1 sites
- Shout out to Joel Snow



Grid Monte Carlo == $$

	Monte Carlo	
Country	Events	$ Equivalent
Brazil	9,353,250	$25,165
Canada	20,953,750	$56,376
Czech Rep	16,180,497	$43,534
Germany	107,338,812	$288,797
India	1,463,100	$3,936
France	106,701,423	$287,081
Netherlands	11,913,740	$32,054
UK	18,901,457	$50,854
US	32,412,732	$87,207
	325,218,761	$875,004

Statistics: 2006



D0 Vital Statistics		
	1997(projections)	2006
Peak (Average) Data Rate(Hz)	50(20)	100(35)
Events Collected	600M/year	2 B
Raw Data Size (kbytes/event)	250	250
Reconstructed Data Size (kbytes/event)	100 (5)	80
User format (kbytes/event)	1	80
Tape storage	280 TB/year	1.6 pb on tape
Tape Reads/writes (weekly)		30TB/7TB
Analysis/cache disk	7TB/year	220 TB
Reconstruction Time (Ghz-sec/event)	2.00	50 (120)
Monte Carlo Chain (GHz-sec/event)	150	240
user analysis times (Ghz-sec/event)	?	1
user analysis weekly reads	?	8B events
Primary Reconstruction farm size (THz)	0.6	2.4 THz
Central Analysis farm size (GHz)	0.6	2.2 THz
Remote resources(GHz)	?	~ 2.5 THz(grid)

Hurray for Moore's law!

Operations 2006-Now

- LHC activities were ramping up
- D0 didn’t stop!
  - we had to find efficiencies
- Focus on Scaling—particularly for SAM
- Focus on Robustness
  - Lazy Man System Administration
  - DB servers Round Robin failovers
- Focus on functionality
  - SAMGrid and interoperability with LCG
- Mike Deisburg and Qizhong Li are the go-to folks!

2014 Statistics

D0 Vital Statistics			
	1997(projections)	2006	2014
Peak (Average) Data Rate(Hz)	50(20)	100(35)	
Events Collected	600M/year	2 B	3.5 B
Raw Data Size (kbytes/event)	250	250	250
Reconstructed Data Size (kbytes/event)	100 (5)	80	
User format (kbytes/event)	1	80	
Tape storage	280 TB/year	1.6 pb on tape	10 pb on tape
Tape Reads/writes (weekly)		30TB/7TB	
Analysis/cache disk	7TB/year	220 TB	1 PB
Reconstruction Time (Ghz-sec/event)	2.00	50 (120)	
Monte Carlo Chain (GHz-sec/event)	150	240	
user analysis times (Ghz-sec/event)	?	1	
user analysis weekly reads	?	8B events	
Primary Reconstruction farm size (THz)	0.6	2.4 THz	50 THz
Central Analysis farm size (GHz)	0.6	2.2 THz	250 THz
Remote resources(GHz)	?	~ 2.5 THz(grid)	~ 0.2 THz(grid)/ year

Thanks!

Gavin: “Wow...where to start :-) - immediate thought - a lot of very good memories....of a lot of hard work from very capable, and fun people :-)”

DØ