FBSNG and Disk Farm - parts of large cluster infrastructure

Large Cluster Computing Workshop

Fermilab

October 21-22, 2002

Igor Mandrichenko, FNAL/CD/ISD

FBSNG – Farm Batch System (Next Generation)

- Developed at FNAL in 1997-98 for Run II production PC farms
- From event parallelism (CPS) to file parallelism (FBS)
- FBS production since 1998
  - was dependent on LSF as scheduler
- In 2000 was redesigned not to use LSF - FBSNG
  - in production since 2000

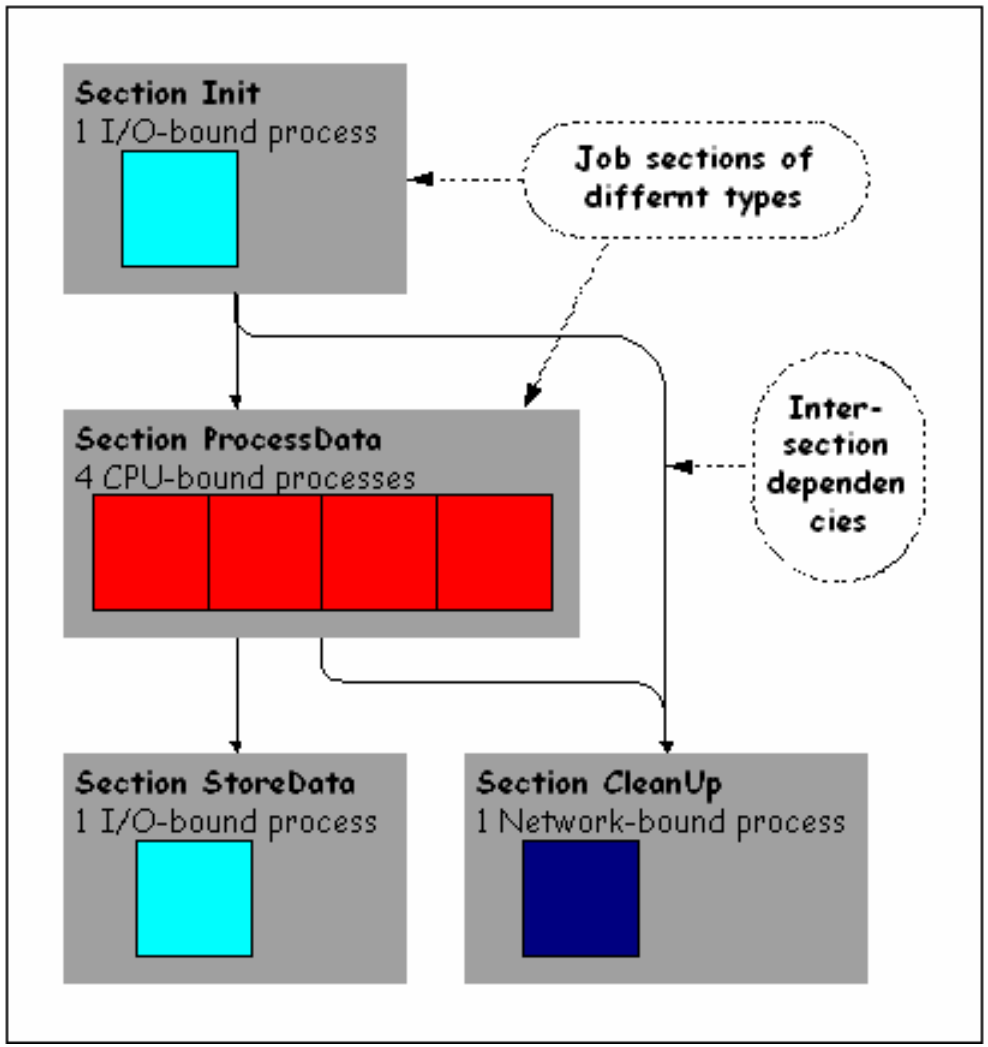
URL: http://www-isd.fnal.gov/fbsng

FBSNG Concepts: Resource Counting

- Instead of load measurement, **resource counting**:
  - Know resource capacity of farm nodes
  - Know process resources requirements
  - Know which process runs where
  - Start new process when and where resources are available

- Makes the system *simple, robust, flexible, portable*

FBSNG Concepts: Job Structure



- Unit of operation is an array of batch processes (*job section*)
- FBSNG job consists of (dependent) sections

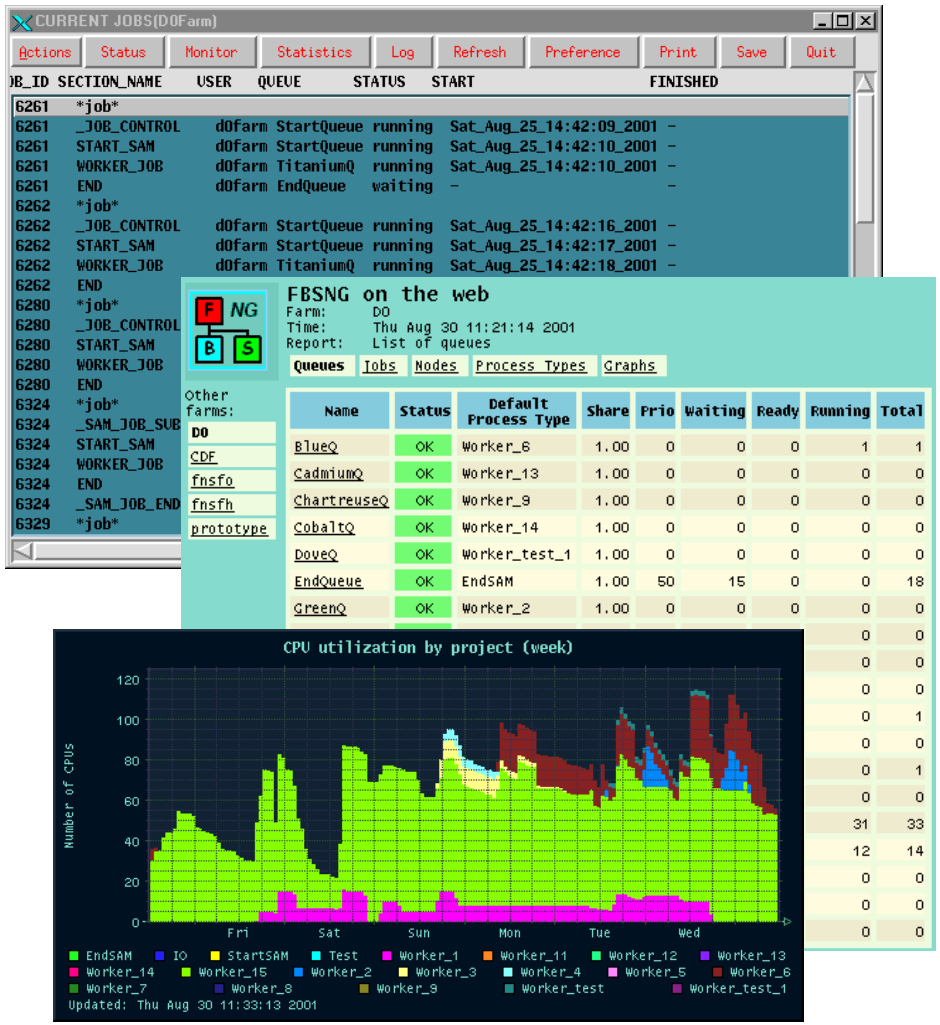
FBSNG Concepts: Abstract Resources

- Abstract Resources
  - All resources in FBSNG are *abstract* counted semaphores
  - **Local resources** – associated and available locally on farm nodes
    - CPU
    - Disk
    - Tape drives
  - **Node attributes** – features of farm nodes
    - OS flavor
    - Installed software
    - Logical attributes (“red”,”green”, used to partition the farm)
  - **Global resources** – resources shared by all the processes on the farm
    - network throughput
    - NFS-exported disks
    - global semaphores
- No predefined resources
- Allows high flexibility in farm/cluster configuration and management

FBSNG Features

- Scheduler
  - Task/project/group prioritization
  - Fair-share scheduling
  - Guaranteed scheduling
  - Resource utilization quotas
- Dynamic farm (re)configuration
- Robustness with respect to failures of individual farm nodes
- Kerberos support
  - Client authentication – ability to access over WAN
  - Creates credentials for batch processes
- Easily portable: supported on Linux, IRIX, SunOS, OSF1
- Recently added Globus Job Manager interface
  - GSI support

FBSNG User Interface



- Command line interface
  - Job submission, control
  - Farm management
- GUI
  - Job monitoring, control
  - Farm management
- Python API
  - Provides full functionality
  - UI, GUI, FBSWWW use API
- Minimal requirements for client-only installation.
  - Access over WAN
- Web interface (FBSWWW)
  - Resource/job monitoring
  - Node status monitoring

FBSNG – Experience

- Currently managed farms:
  - CDF, D0 on-line production farms (150+ computers each, growing)
  - “Fixed Target” (common use) farm (106+2 computers)
  - CMS USA Tier 1 center (3 farms ? - see Hanses talk)
  - CAF - CDF Analysis Farm
  - NIKHEF
  - Other HEP sites, one known corporate site

- FBSNG is full scale batch system for farms and clusters
  - Has worked well on farms of different sizes with various kinds of users and resource utilization patterns
  - Robust, low maintenance, easy to deploy, support and manage

Disk Farm - Distributed Data Storage

- Typical computing farm can be viewed as an array of disks controlled by an array of CPUs, or **disk farm**:
  - Capacity: 100 nodes * 2 * 30 GB = 6TB
  - Throughput: 10 MB/s / node * 100 nodes = 1 GB/s
- Price: ~100GB disk for $100 -> $1/GB
- Typically, each 30 GB disk is managed by 1GHz CPU
  - Compare to ~1TB and 8 300MHz CPUs (d0bbin)

- BUT, utilization is difficult:
  - Highly distributed storage
  - Unreliable components
  - Access and allocation must be coordinated

- Disk Farm is a product which helps utilize large unused disk capacity of farm nodes

Disk Farm – Distributed Data Storage

- Organizes distributed disk space spread over nodes of the farm into global *virtual* name space
  - Virtual file path: /e123/runII/data/mc123.dat
  - Physical file path: node1:/local/stage1/dfarm/xyz123

- User interface:
  - get, put, mkdir, rmdir, rm, ls commands – similar to UNIX FS access commands. E.g.:

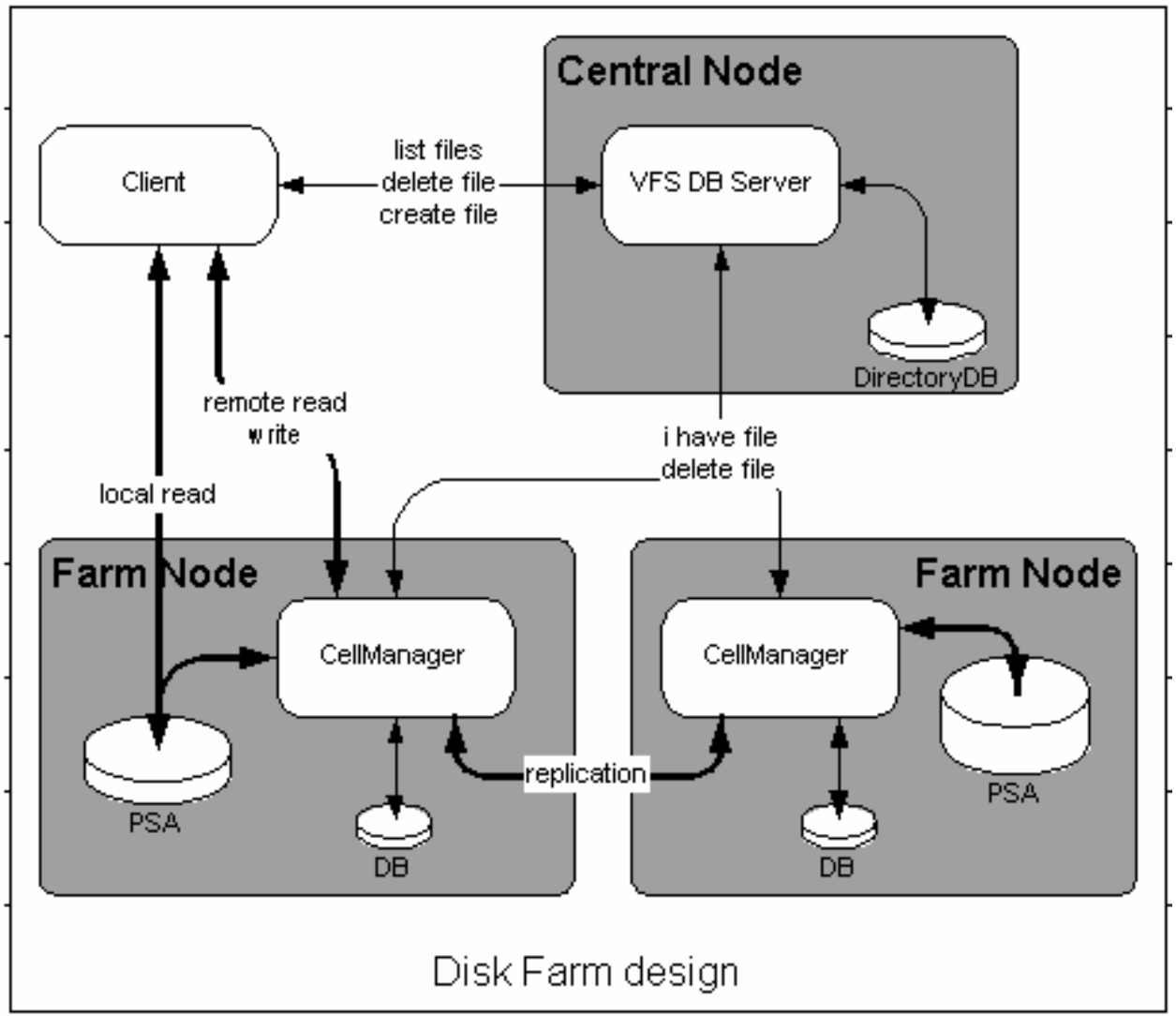
```
$ dfarm mkdir /e123/runII/data
$ dfarm put /scratch/mc123.dat /e123/runII/data
$ dfarm ls /e123/runII/data
```

  - On the node where the data happens to be:
    - "get" is local – faster, cheaper
    - Ability to read data without copying out of disk farm
    - "put" is amlost always local
  - Users are limited by (optional) global quotas, not physical sizes of individual volumes
  - POSIX semantics access being considered

Disk Farm Features

- Data replication:
  - User: here is my file, make 3 copies of it on 3 different nodes
  - In case 2 nodes go down, 1 copy is still available
  - Replication is performed off-line, without user waiting
  - Data can be re-replicated later to maintain desired number of replicas
  - If a node is to go down, its contents can be replicated
- Load management and balancing
  - Automatically chooses one of least busiest nodes
  - Configurable limits on number of active transfers per node
  - Nodes can be made read-only (put on hold)
- Scalability
  - Scales naturally with the farm size
- WAN access
  - Kerberized FTP server
  - Work on GSI GridFTP interface

Disk Farm Design



Disk Farm - Experience and Status

- Robust, reliable, low maintenance product
- Best if used as a temporary storage of data on the farm
- Current installations:
  - CFD production farm 6.7 TB on 170 nodes
  - D0 farm: 2.1 TB on 180 nodes
  - Fixed Target farm: 1.8 TB on 90 nodes

URL: http://www-isd.fnal.gov/dfarm

Farm Resources Management Tools

- FBSNG and Disk Farm are farms infrastructure tools used to organize and coordinate use of resources provided by computing farms:
  - CPU power
  - Local disk
  - Inter-node network bandwidth
- They provide local and, trough Grid interface, remote access to computational resources