

CDF Production Farms in 2002

Stephen Wolbers, Fermilab
For the Farms Group
December 6, 2001

Outline

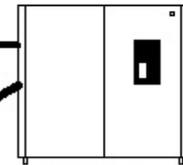
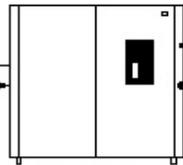
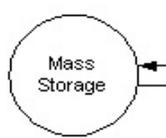
- Farms Hardware and Capabilities
- ENSTORE changes
- Reprocessing from PADS/secondary datasets
- Processing Ideas

Status of CDF Farms Hardware

- **137 PC's are in place.**
 - 50 PIII/500 duals (couple of years old)
 - 23 PIII/800 duals (1 year old)
 - 64 PIII/1 GHz duals (recent arrivals)
 - ▶ Equivalent to 430 PIII/500 CPU's
 - ▶ Compare to 375 estimated need for full-rate Run 2
 - Can add more if necessary
- **2 I/O nodes, both SGI O2000**
 - fcdfsgi1 (shared with data logging)
 - 2 Tbyte of disk for I/O
 - cdfarm1
 - 1 Tbyte of disk for concatenation

Run II CDF PC Farm

11 MB/s



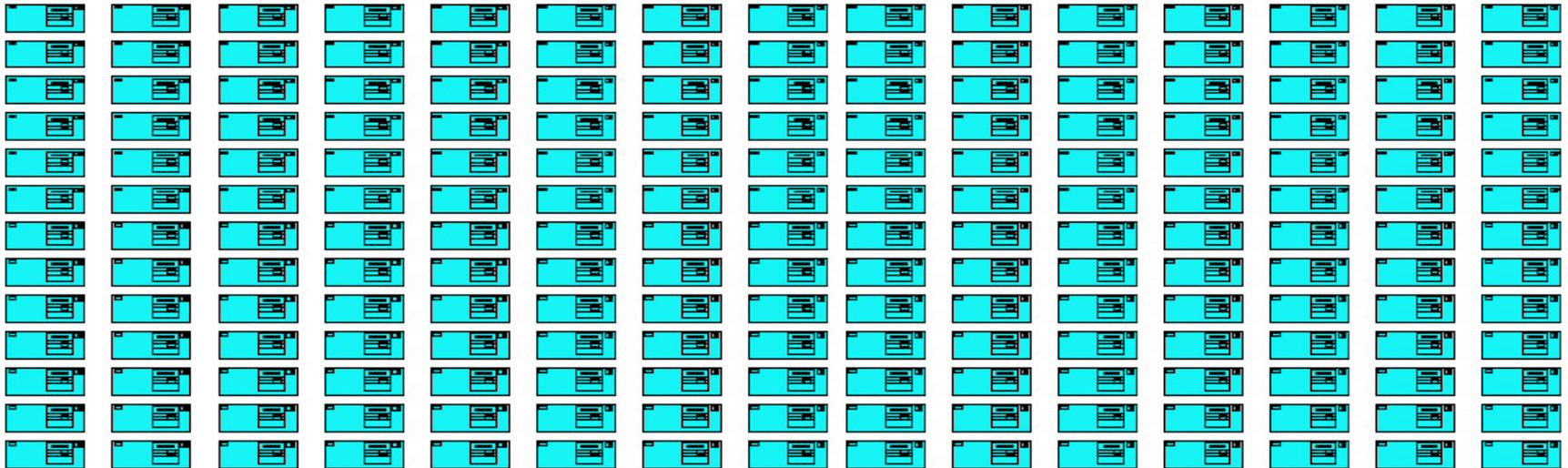
Network Switch

20 MB/s

5 MB/s

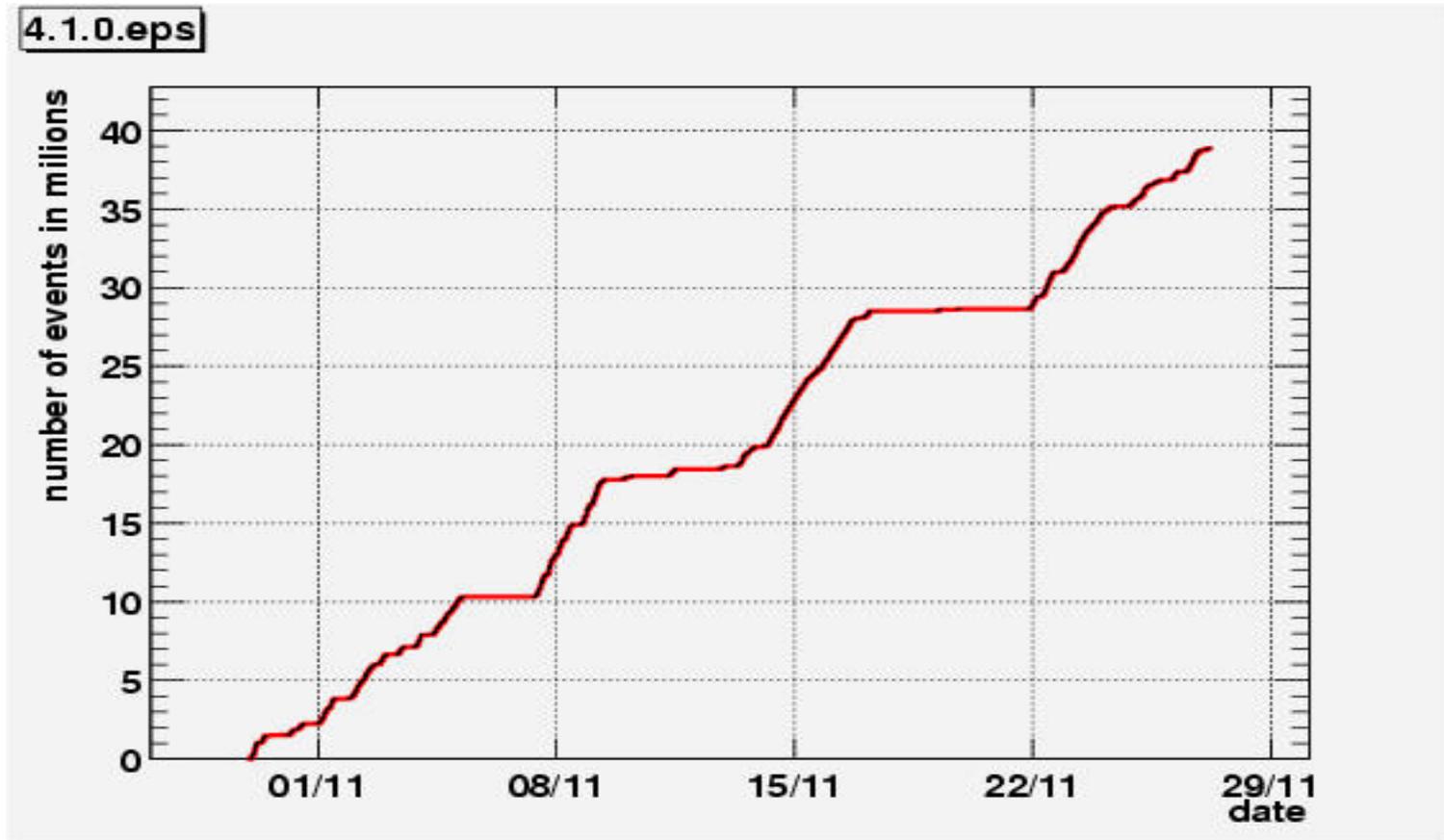
fcdsg1

cdfarm1



Farms Performance at present

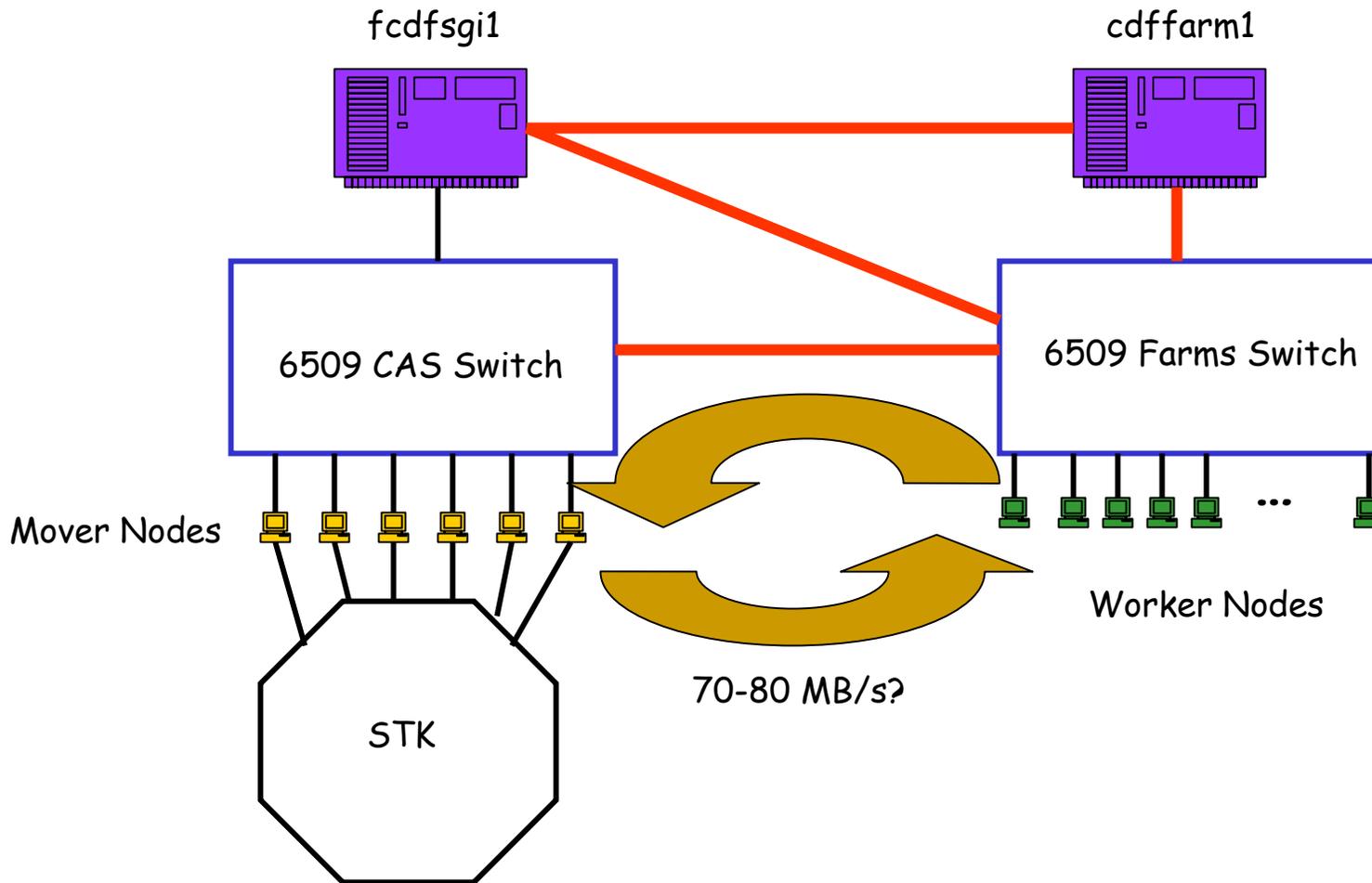
- Can Run Approximately 1-2 million events/day Using Current DH/tape system



ENSTORE Changes

- Farms/DH will change the I/O system by using ENSTORE as the underlying tape access and moving to STK 9940 tapedrives.
- The details are not all worked out yet.
 - Farms will likely read/write directly to tape using encp.
 - This reduces latency for two reasons:
 - Can write to tapes with < 1 tape's worth of data (Plan to use 1 fileset (10 GB))
 - Tape write speed is 10 MB/s instead of 3 MB/s
 - Thinking about how to reduce latency even more by read/write to disk cache and not wait for tape.
 - If we move concatenation to the worker nodes, another bottleneck is removed.
- **Goal: Scalable system that can handle multiple streams simultaneously.**

Enstore/Farms (Proposed)



Reprocessing from PADS and Secondary Sets

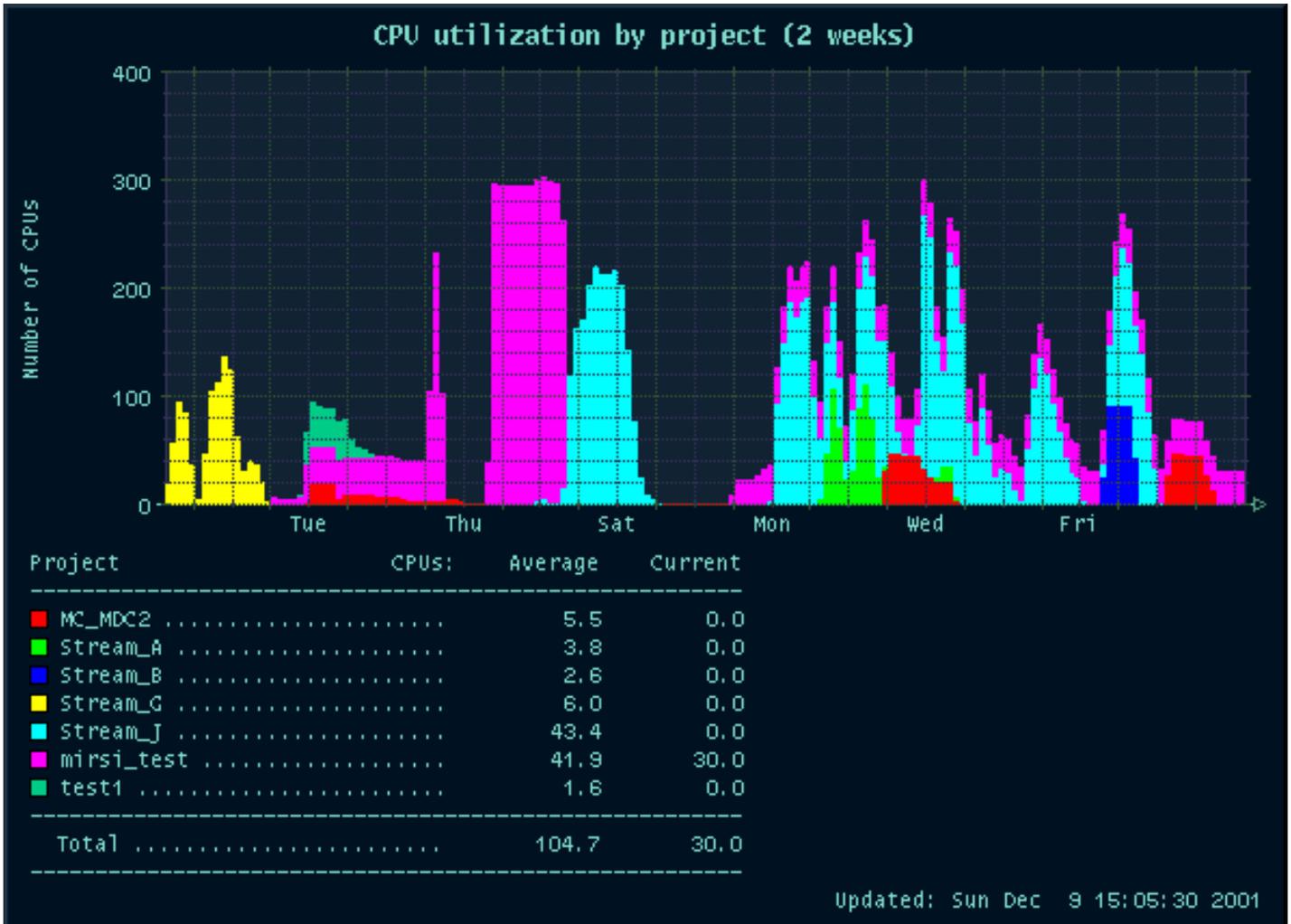
- **Requirement:**
 - ProductionExe has to be able to correctly process with farms output or secondary datasets as input.
 - Calibrations must be available and up-to-date.
 - CPU/event doesn't really matter much.
 - Size of files should be close to 1 Gbyte if possible
- **To run on the farms without major changes:**
 - PADs and/or 2ndary sets are on tape in DH system
 - Reprocessing is "just another stream"
- **Major changes required if:**
 - If sets are not in DH system but are just a "set of files" on some computer.
 - This doesn't seem like a good idea anyway.

Processing Ideas

- Superexpress will continue to run on Stream A on fcdfsgi2.
- Processing of newest data will occur using “frozen” ProductionExe.
- Reprocessing can occur in parallel.
 - The farm is logically subdivided into farmlets.
- Monte Carlo will run whenever files are ready.

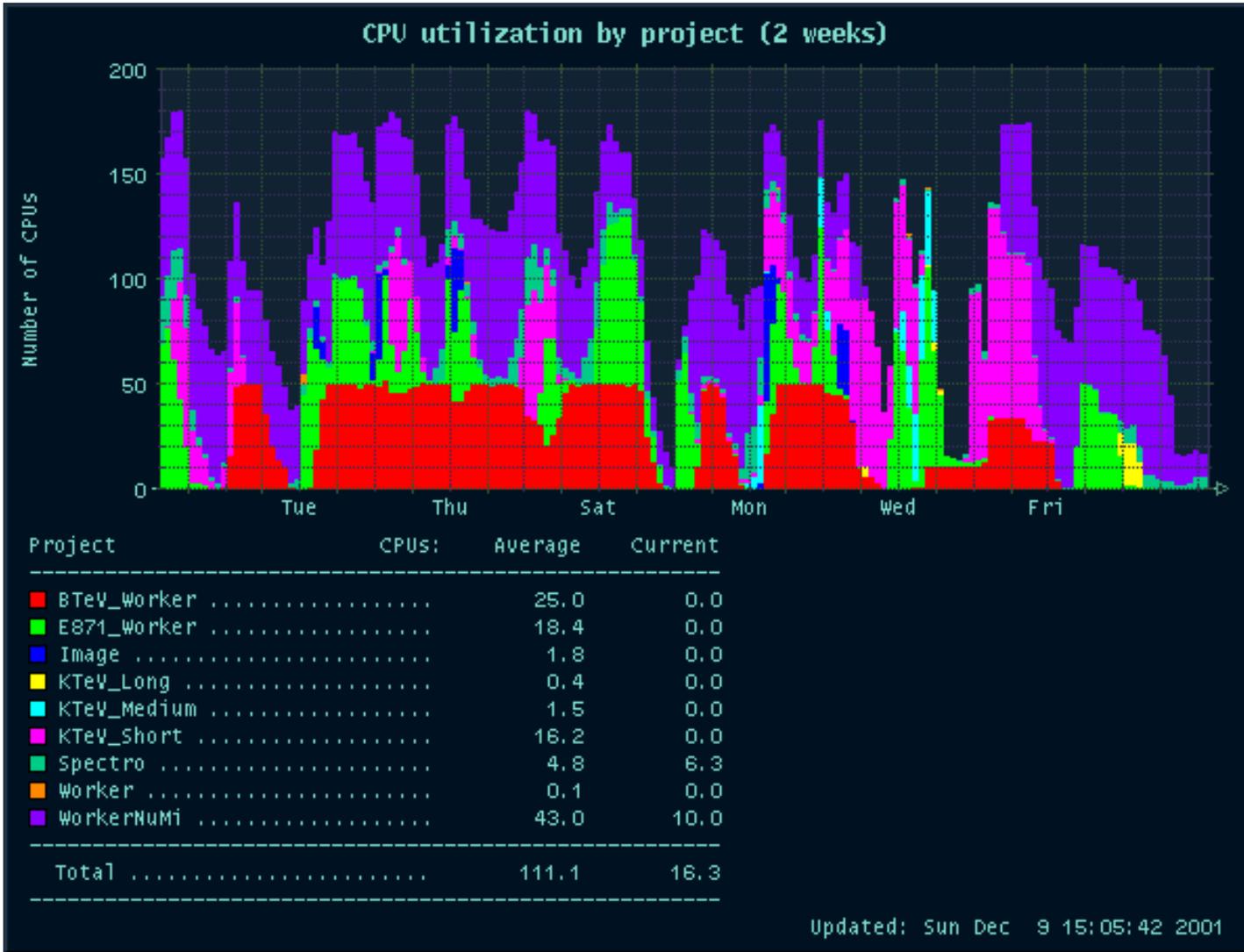
Optimal Farm Operation?

- To optimize throughput, the farm processes a small number of streams.
 - Easier to balance input/output.
 - Smaller chance of “deadlocks” due to missing input, too many output streams, filling output disk while waiting for tapedrives, etc.
- Slowly increasing the mix of jobs.
 - 2 streams + Monte Carlo.
 - Will add new data soon.



Enstore/ STK 9940 operation

- Working on ideas that give vastly increased input/output and parallel operation.
- DO is starting to use the system - experience is very good so far.
- Rest of Fermilab Farms has been using for quite some time.
 - Many independent users
 - Complicated priority scheme
 - It works



Some scenarios

- **No change to current system:**
 - 2 million events/day in 2 streams is certainly achievable.
 - 7 MB/s is about 28 Hz or 2.4 million events/day.
 - Rotate among streams.
 - More streams will likely overstress the input and/or output systems on fcdfsi1.
- **More disk on fcdfsi1**
 - Still 2 million events/day (tape drive limitation).
 - Probably can handle >2 streams at a time.
- **Enstore I/O**
 - Potential for a big improvement if data can be streamed straight into the farm at the rates expected.