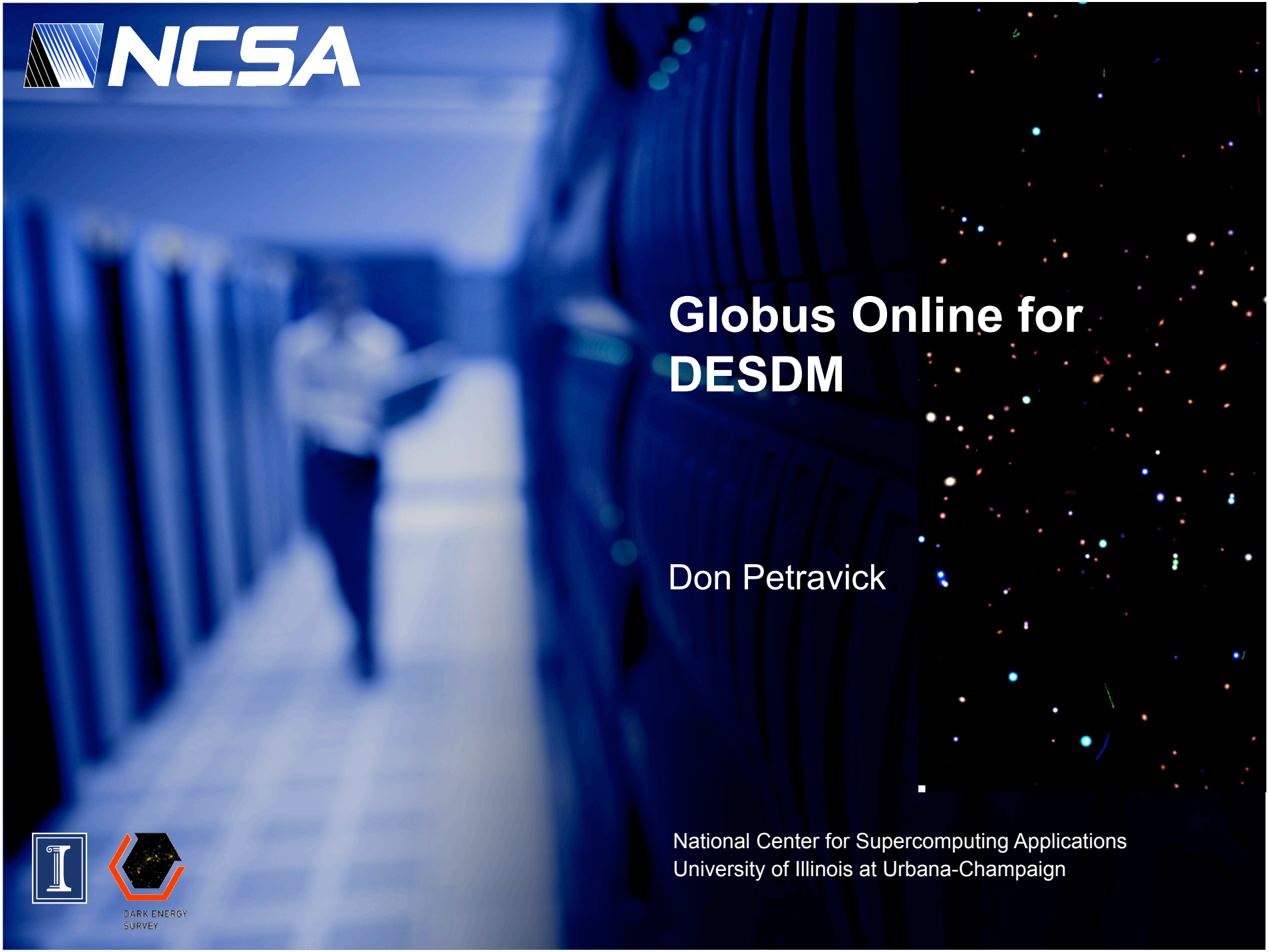


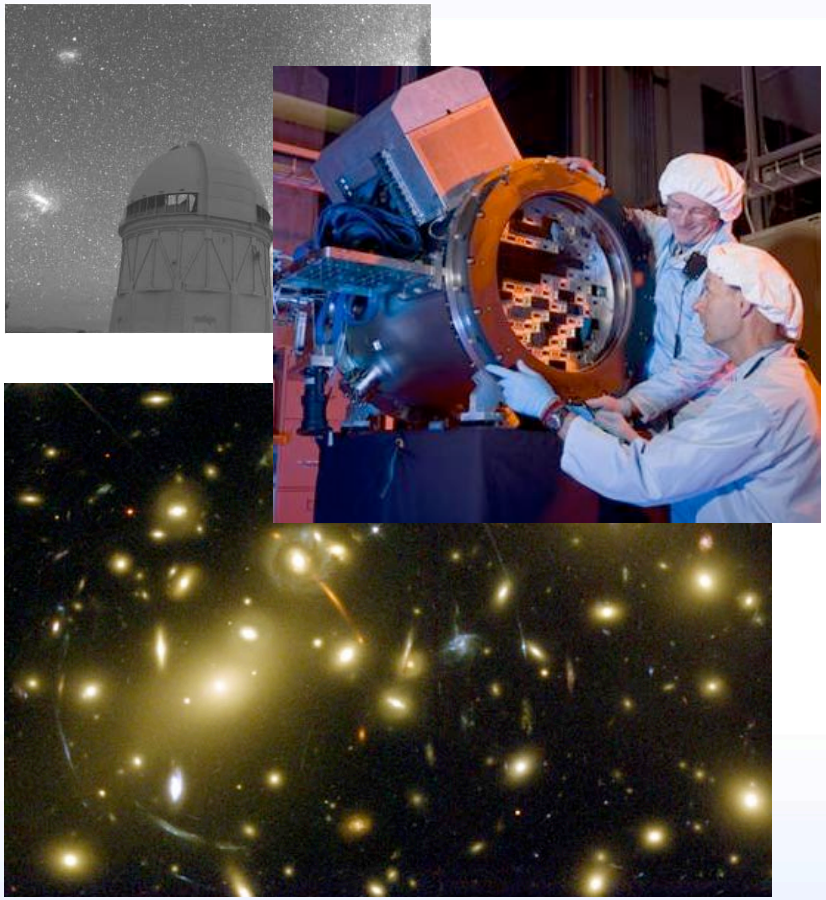
Globus Online for DESDM

Don Petravick

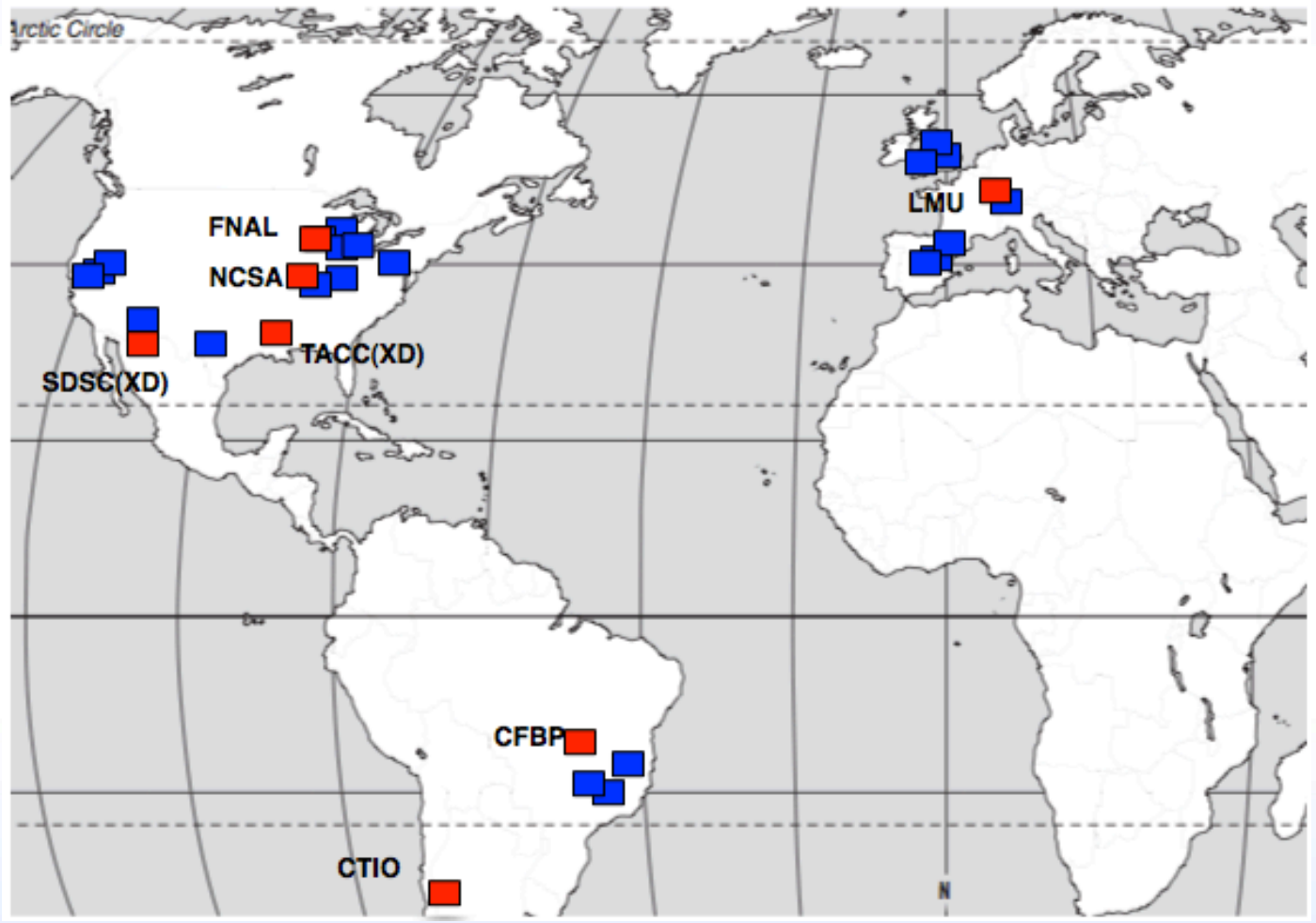
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

Dark Energy Survey

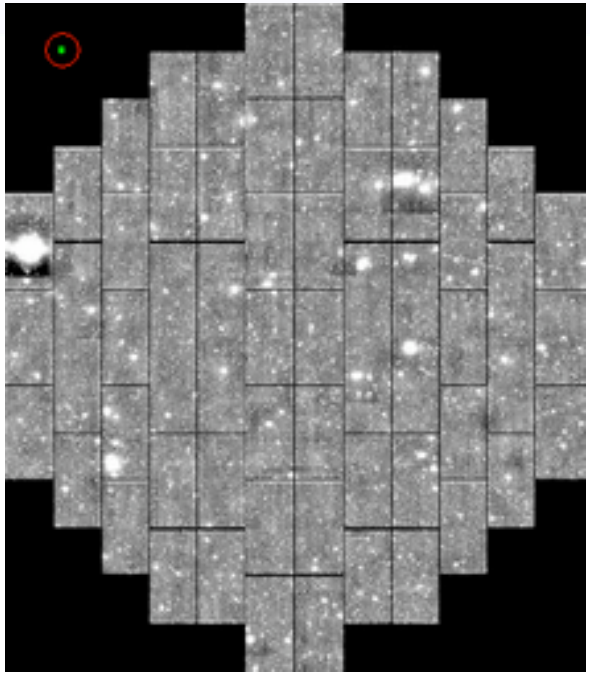
- DES combines four probes of Dark Energy:
  - Type Ia Supernovae (SN)
  - Baryon Acoustic Oscillations (BAO)
  - Galaxy clusters (GC)
  - Weak Gravitational Lensing (WL)
- Program
  - Five years,
  - 525 nights of observation
  - 300 million galaxies that are
  - Survey will image 5000 square degrees
  - 5 optical filters to obtain crude spectra
- Also observe smaller patches to discover and study thousands of supernovae.

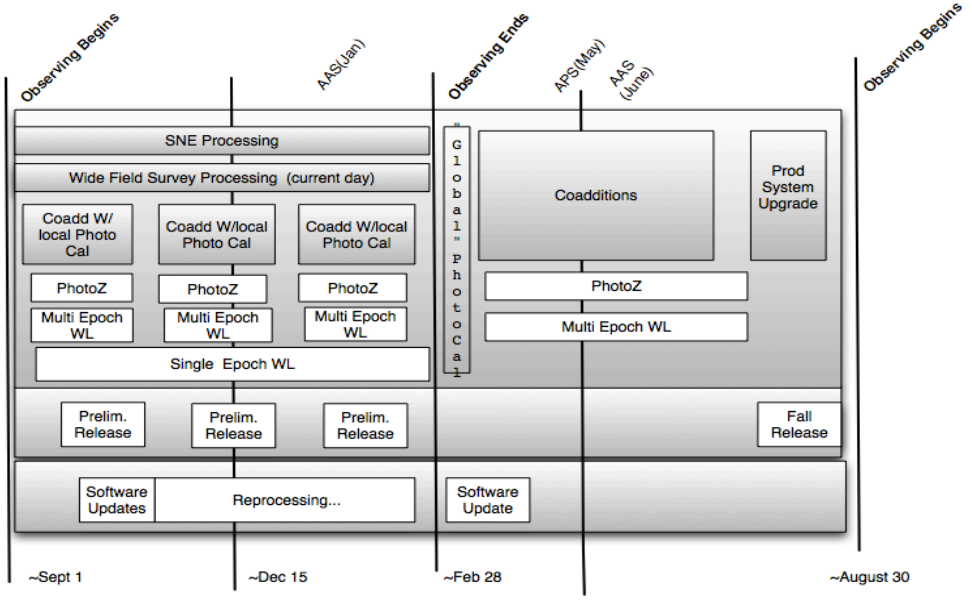


Scientists and Science Facilities



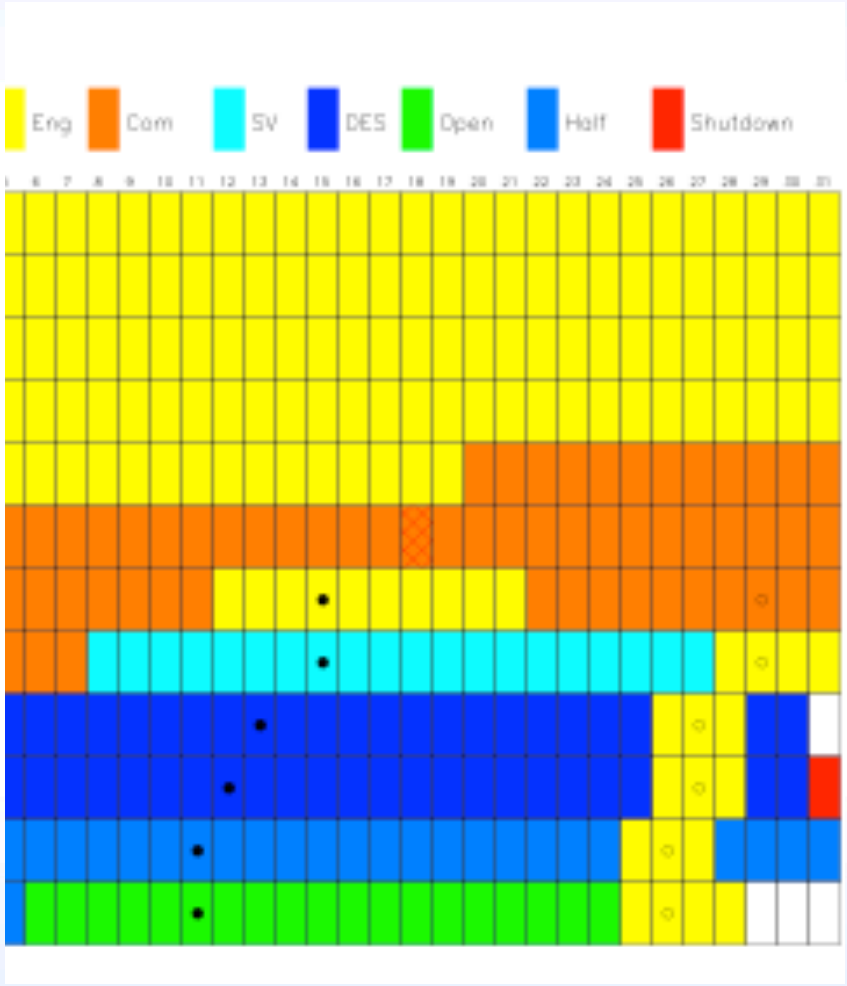
Files: Cant live with them, can't live without them…





- DESDM is made up of many state of the art codes.
  - Few are be-spoken -- Composed (semi) autonomously
  - PI's involved in a continuum of projects,
- Standards world of astronomy => files

Demanding Schedule

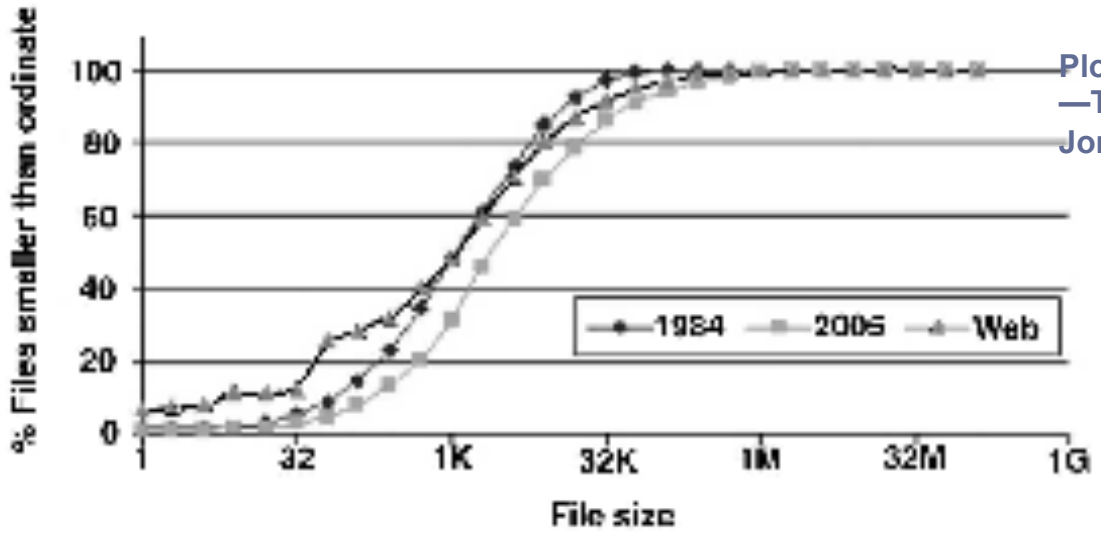


- While on-sky, we need to keep up with certain processing.
  - 80% of the time, the survey table will be updated astronomical sunset at CTIO on the following night.
  - 80% of the time, SN fields will be processed by sunset at CTIO on the following night.
  - Otherwise the processing of the wide field data shall be completed within 3 business days. 80% of the time.
- XSEDE -> Processing capability on 3 machines.
- Implies transport anywhere.

NSF environment for "big small projects"

- OK, we're not an MREFC.
  - Look for leverage, not silos
- "Many research projects require access to computational, data storage or visualization resources in order to complete the work proposed. For those projects that require such resources at a scale that is beyond that typically available locally, NSF provides the TeraGrid. …."
- DESDM:
  - Database, central file store: Urbana, IL
  - Bulk Compute:
    - UCSD (~2000 mi)
    - TACC (~1000 mi)
    - DR – FNAL (160 mi) + ???

File Sizes do not want to grow exponentially



Plot: File Size Distribution on UNIX Systems —Then and Now: Andrew S. Tanenbaum, Jorrit N. Herder*, Herbert Bos

- DESDM:
  - > 2PB of (compressed) data files.
  - Median file size …. Few 100 KB.
- Pattern has been observed at DOE and other HPC centers to this day.

Basic unit of “focus “ in DES

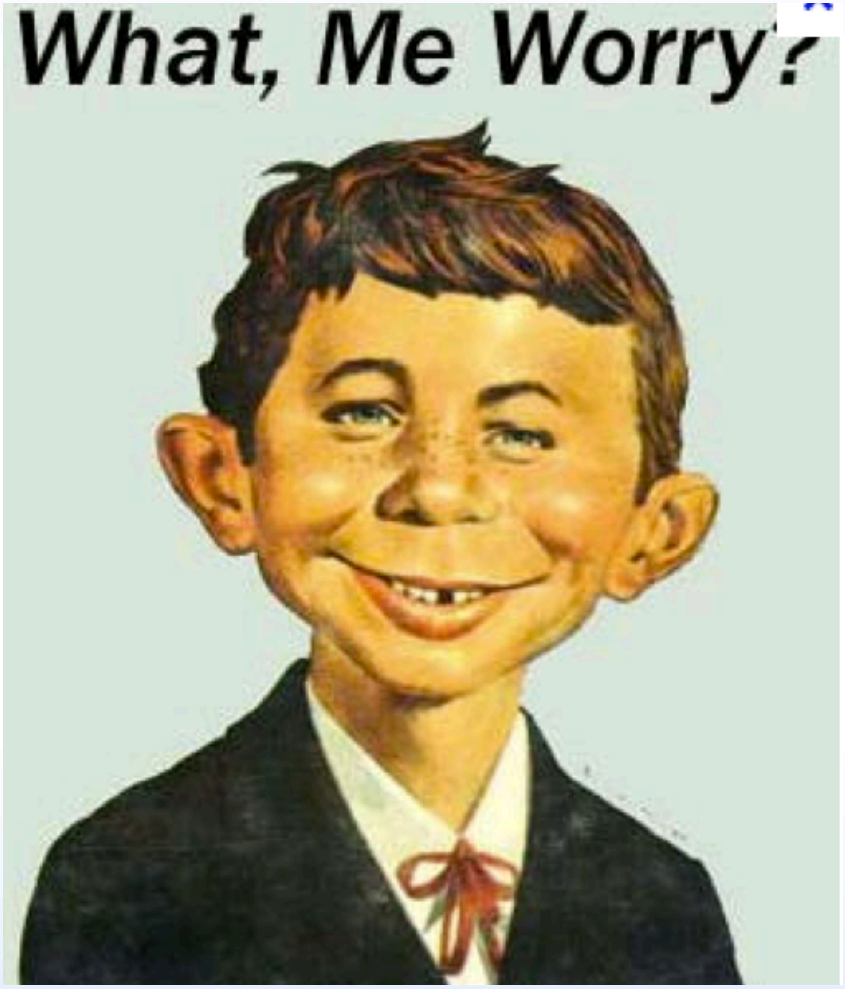
- Information
  - CCD: 1024 x 2048 X 16 bit == 4MB.
  - Exposure: 62 x 4 MB == 248 MB
  - Tile: 0.75 sq degrees. (many per “exposure”) ~ 50 MB.
  - Most “sky” is dark -> (lossy) compression x5, lossless: x2.
  - Calibration and other files << above files.
- Median “natural: file size
  - is small
  - Will not grow over lifetime of system.

So, all is well, right?

- Many small files.
- Codes you don't control.
- Bulk compute >>

1000 km away?

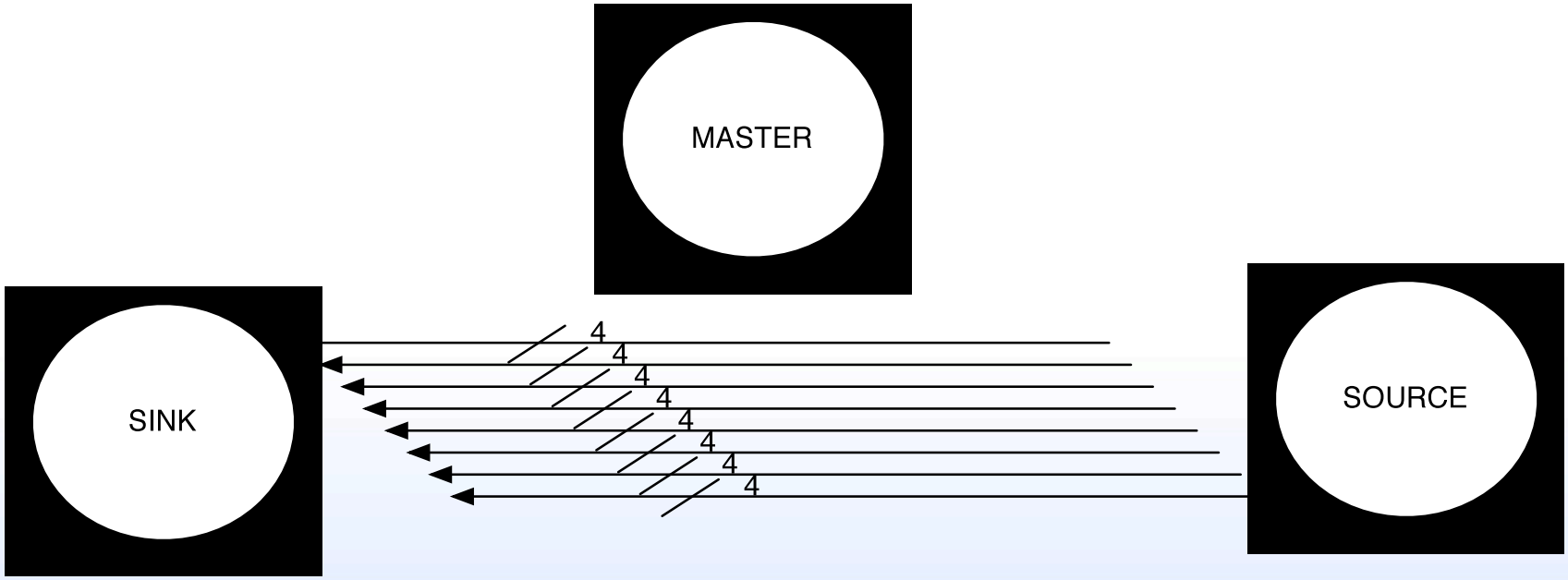
- Lean budgets, pressure to do more with less.
- Not problem not understood by many.



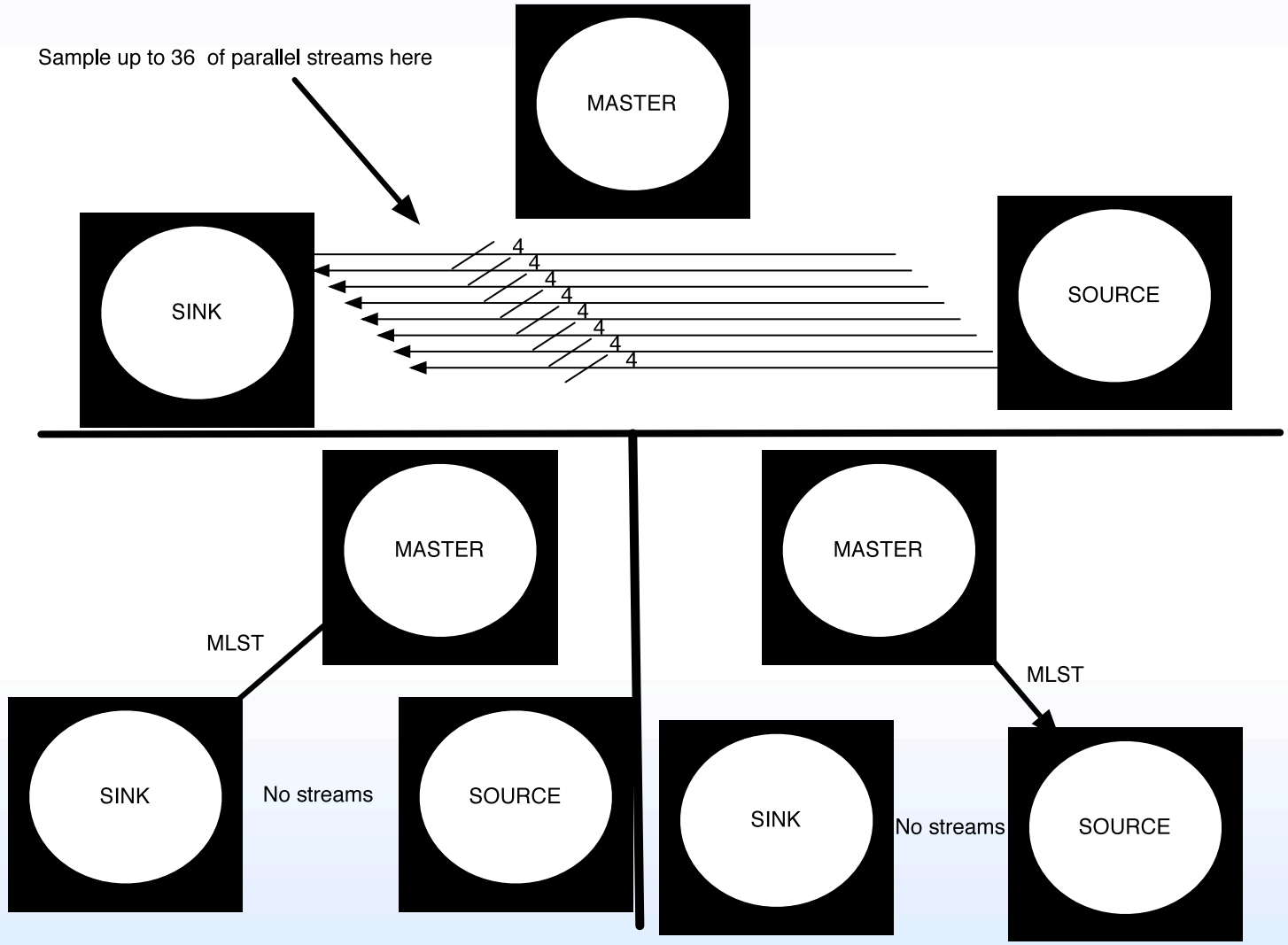
Programmer Thought…

After some in-field measurements:

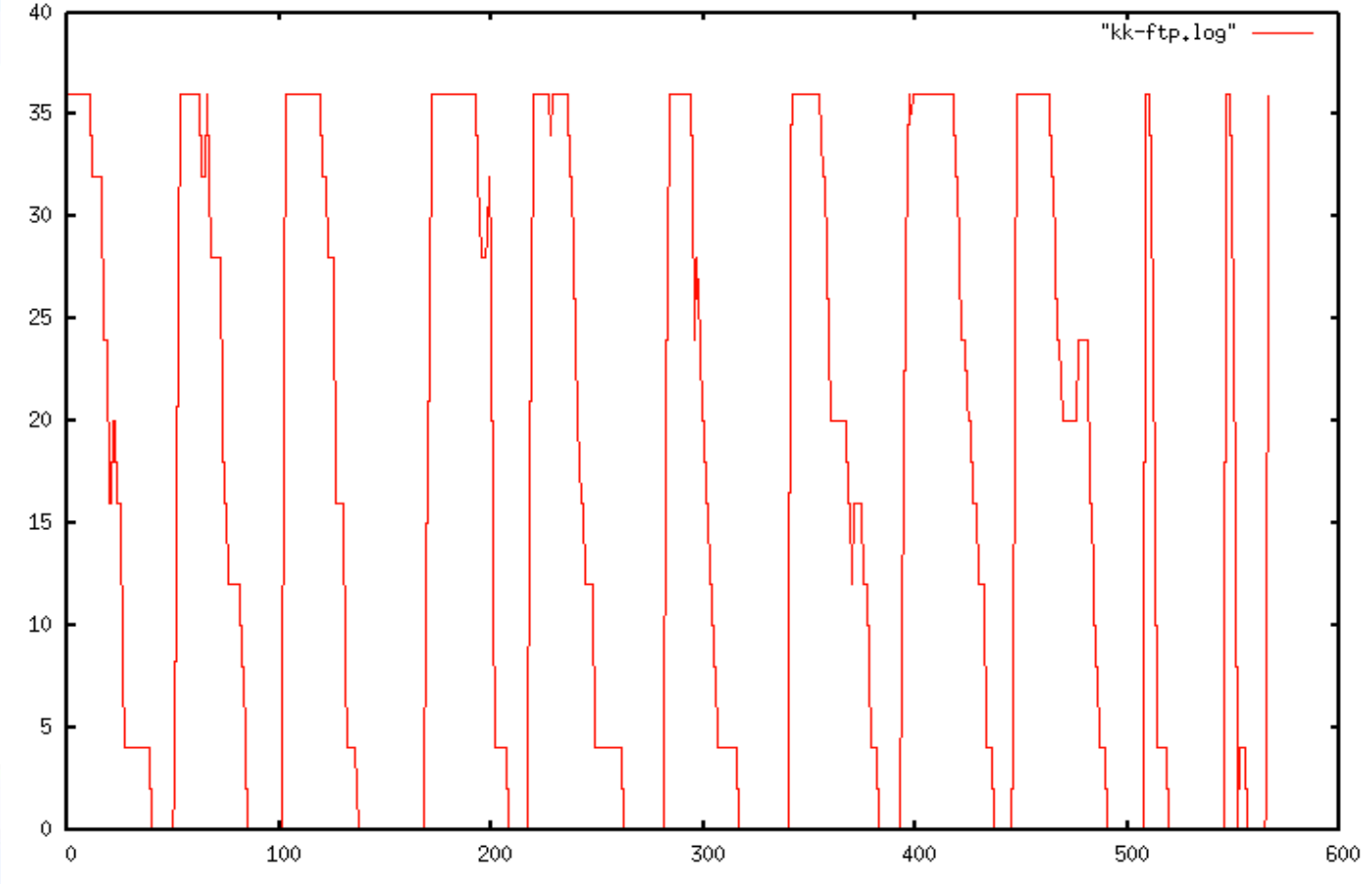
- Divide xfer into batches of k files (K now lost to history)
- Each unit of xfer is 1/9th of the files.
- Transfer each file in stripes of 4.
- Check source against “truth” before and after xfer.
  - But checks “must” be trivial compared to xfers.



So observe # open TCP streams ….



Issues to resolve in file transfer performance



Open Transfer Sockets v.s Time (seconds)
Transfers from NCSA-> TACC (35 ms latency)

Observations

- Technical
  - > 50% dead – cause: assumptions in mind of programmer.
    - False assumption -- “Verify” a trivial amount of time.
    - Dividing work ahead of time in a slice by the number of files.
    - Wait for all nine to complete before beginning check
  - MLST: request-response MEP is linearly worse w/ latency.
    - Move from local (e.g. “abe”) to remote
  - Also– a lot of mechanics -- under the sheets:
    - A mix of uberFTP + globusURL copy.
    - Perl gymnastics to write, parse command files, pu tthem all into /tmp but provide for debugging
    - (guess) Programmer focused on mechanics.

Basic lessons surfacing

- Efficiency follows specialization. (Smith, 1776)
  - Lemma : Its even more efficient for specialized people to build smarts into tools, not people. (Hayek? 1945?)
- You can have any color you want as long as it's black(Ford, ~1915).
  - Exponential increase in Capabilities (Moore, 1965).
  - Cost per pound for commodity goods is constant (GM, 1935)
    - Capacity compute in storage is cheap.
  - The number of files will explode, given exponential increase in storage… (Tannenbaum, 198x).
- 90% of …network… code….[deals with] Errors (Gosling, 2003)

"A business absolutely devoted to service will have only one worry about profits. They will be embarrassingly large." (Ford ~1920)

Observations (meta)

- Experiment:
  - Responds by (further) running this code in parallel.
    - “obvious” way to cut time.
    - Gets GridFTP error when (3 or 4 or 5 or 10) * 36 streams are active.
    - Blame “darn grid FTP servers”
      - or whatever strikes them….
  - Ideally, no incentives for abuse.
    - But this requires iteration, so that technical can converge to ideal.
    - I can’t afford that for non-astronomy codes.

Conclusions

- Astronomy/ DES benefits from “renting” focused task-oriented specializations.
- Lean budgets make make it impossible for DES to achieve specialized knowledge within its project.
  - For many tasks, I have
    - The kind of people who can learn these things….
    - … not the kind of people who have done these things.
  - Need to focus on astronomy domain.
    - Not in solving and perfecting infrastructure
- (Obvious) other places principle can be applied
  - AAA
  - Storage
  - RDB