Building a Real Workflow

Thursday morning, 9:00 am

Greg Thain <gthain@cs.wisc.edu>

University of Wisconsin - Madison

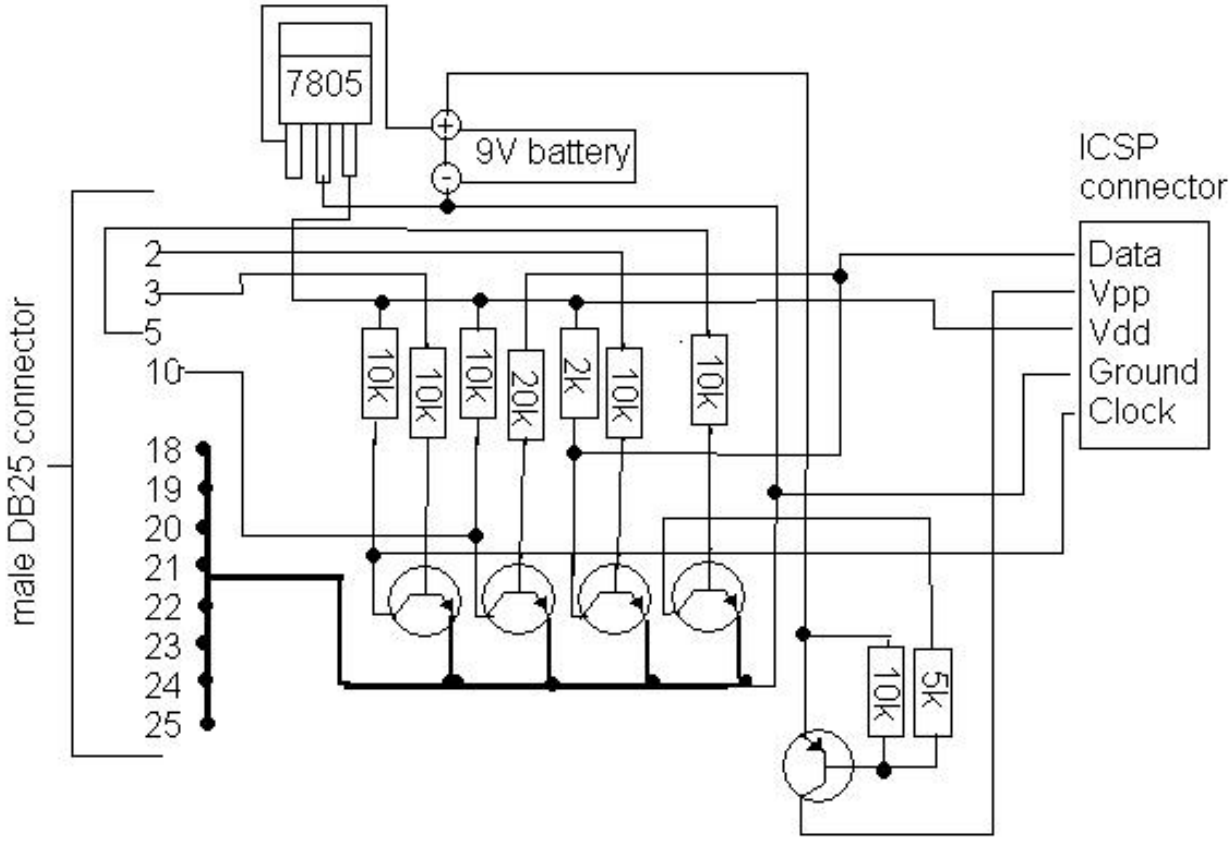
Open Science Grid

Overview

- From script…
  - Pragmatics
  - Estimation
- To production

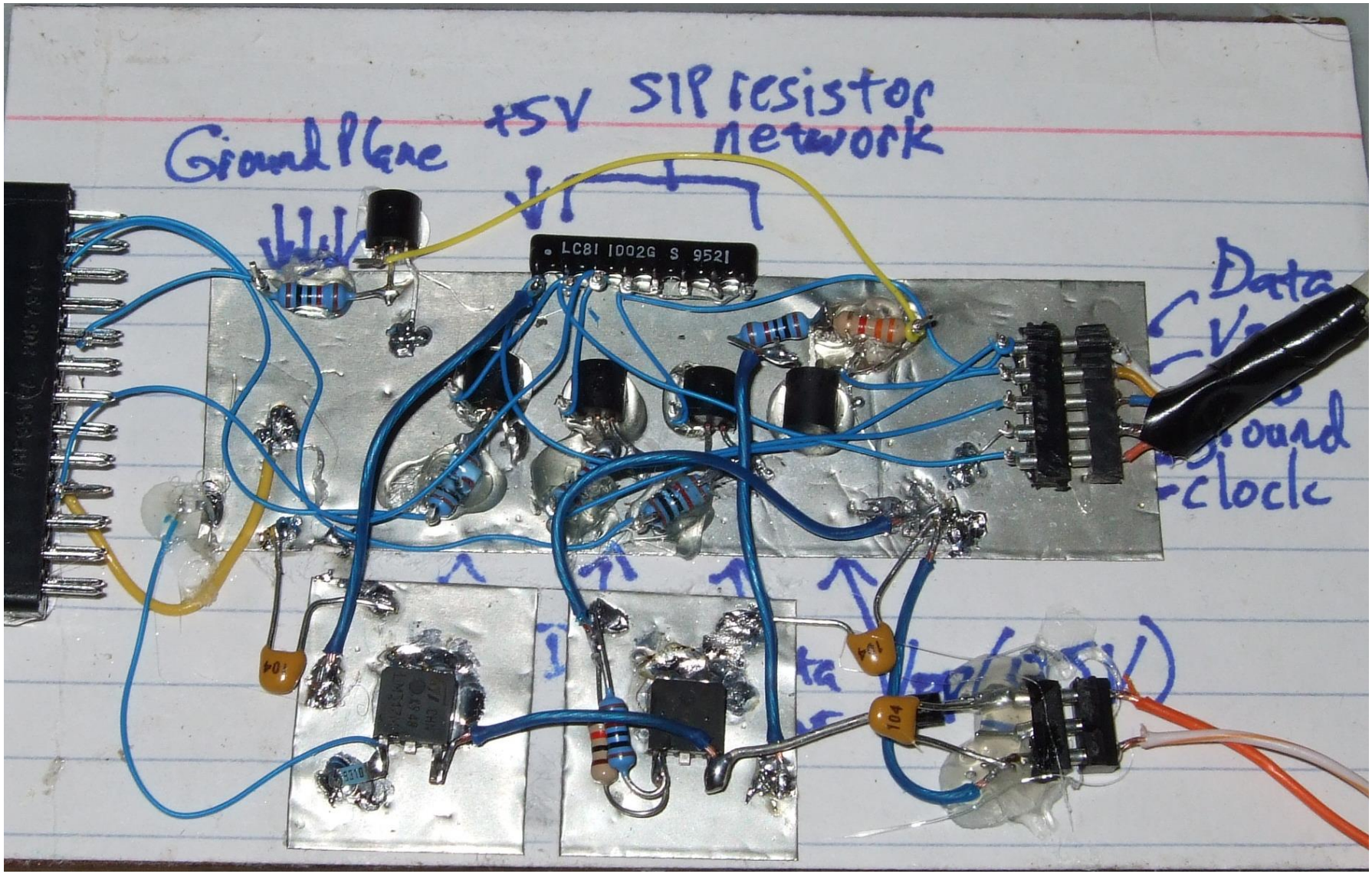
Open Science Grid

From schematics…



Open Science Grid

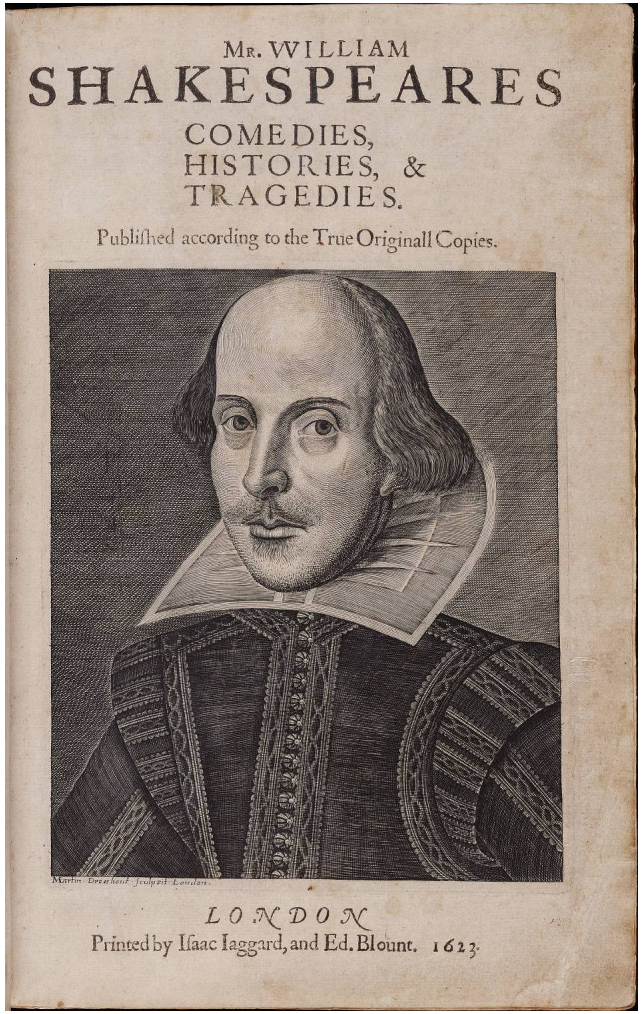
… to the real world



It starts with a script..

```
#!/bin/sh
# Comment

while read n
do
   md5sum $n
done > output
```



Scriptify as much as possible

- What is the minimal number of manual steps?
- Even 1 might be too many.
- Zero is perfect!

First run locally: To measure usage

- Did you get all the inputs?
- Did it run correctly?
  - Are you sure?
- Run once remotely
  - NOT ON SUBMIT MACHINE!
  - Might be surprised
- Once working, run a couple of times
- If big variance, should you take the…
  - Average? Median? Worst case?

Open Science Grid

Estimations: Orders of Magnitude

- Don't sweat the small stuff
  - AMD == Intel == 2.4 Ghz == 3.6 Ghz
- But pay attention to the big stuff:
- What makes this hard?
  - From nanoseconds to terabytes
- Powers of Ten, by the Eames

Resources Jobs Need

- CPU
  - Wall Clock vs. CPU cycles
- Disk
  - Working (execute side)
  - Total (submit side)
  - Bandwidth
    - File transfer queue time
- Network bandwidth
  - Usually for file transfer only

User Log shows all

```
005 (2576205.000.000) 06/07 14:12:55 Job terminated.
        (1) Normal termination (return value 0)
                Usr 0 00:00:00, Sys 0 00:00:00  -  Run Remote Usage
                Usr 0 00:00:00, Sys 0 00:00:00  -  Run Local Usage
                Usr 0 00:00:00, Sys 0 00:00:00  -  Total RemoteUsage
                Usr 0 00:00:00, Sys 0 00:00:00  -  Total Local Usage
        5  -  Run Bytes Sent By Job
        104857640  -  Run Bytes Received By Job
        5  -  Total Bytes Sent By Job
        104857640  -  Total Bytes Received By Job
        Partitionable Resources :    Usage  Request
           Cpus                 :                1
           Disk (KB)            :   125000   125000
           Memory (MB)          :       30      100
```

High Throughput

- What Isn't High Throughput?
  - Quick starting jobs
  - Very short job runtimes
  - Micro optimizations
- What is?
  - Constant job pressure
  - Many jobs
  - Long total workflow times

Rules of Thumb for OSG Jobs

- CPU (single-threaded)
  - Use between 5 minutes and 2 hours wall
    - Upper limit somewhat soft
- Disk
  - Keep scratch working space < 20 Gb
  - Minimize OSG_APP_DATA usage
  - Submit disk usage depends on machine
  - Intermediate needs vs Sinks/Sources

Rules of Thumb (cont.)

- Network
  - Primarily file I/O
  - Fill in FILE SIZE / RUN TIME HERE
- Use squid caching where appropriate

How to apply rules of Thumb

- Not always clear cut
- Need to keep rules of thumb in mind

- May need to join many short processes
  - Easy with shell wrapper
  - But careful with error conditions

- Or break up one big one…

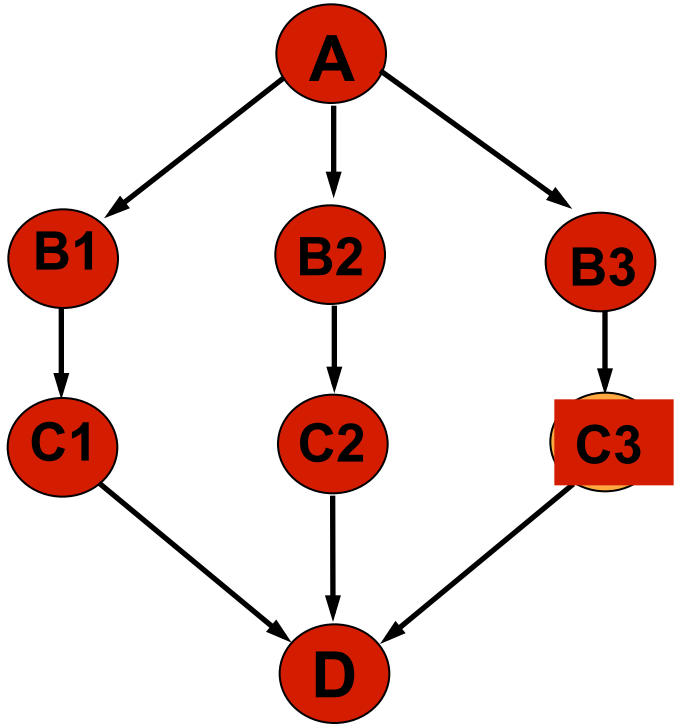
Golden Rules for DAGs

- Beware of the shish kabob!
- Use a single user log
- Use PRE and POST script
- RETRY is your friend
- DAGs of DAGs are good

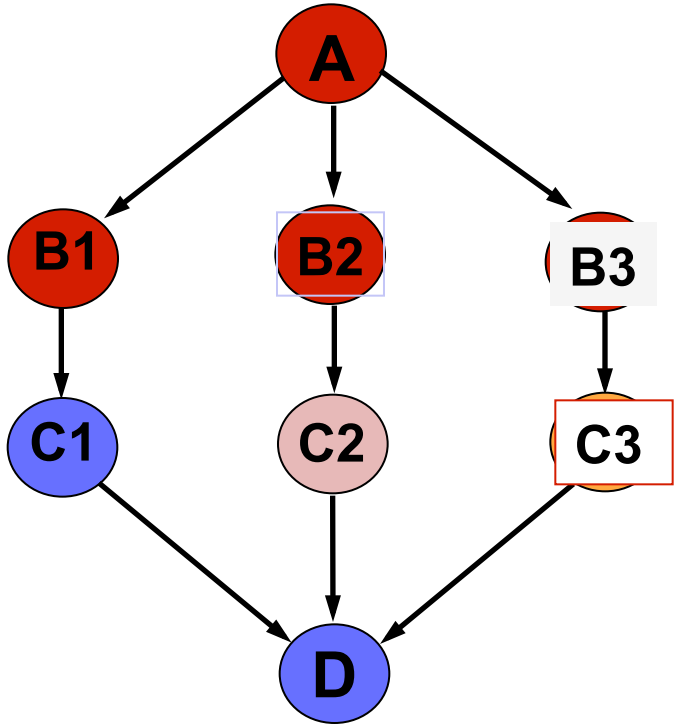
Do you have the full DAG

- Are there manual steps (esp. at beginning and end) that could be automated?

Start with "Conceptual DAG"



Map to Actual workflow based on resource usage



Open Science Grid

Two transforms

- Merging nodes
- Splitting nodes

Open Science Grid

Merging is easy

- Scripting
  - Avoids transfer of intermediate files
  - Careful with error conditions
  - Debugging can be a bit tricky

Breaking up is hard to do…

Ideally into parallel jobs

not always possible

Often need to checkpoint

Standard universe can help

User-defined checkpointing

Checkpoint images can be hard to manage

Putting it all together

- Should have one functional dag
- With appropriate run times

Open Science Grid

Questions?

- Questions? Comments?
  - Feel free to ask me questions later:

    Upcoming sessions