THE SUPERCOMPUTER COMPANY

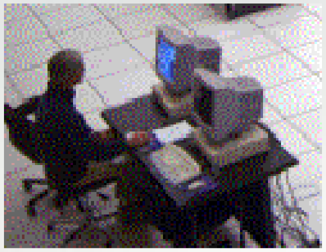
Running on XT Compute Nodes

Kevin Roy

CSC, Finland

September 21-24, 2009

Job Launch – The process

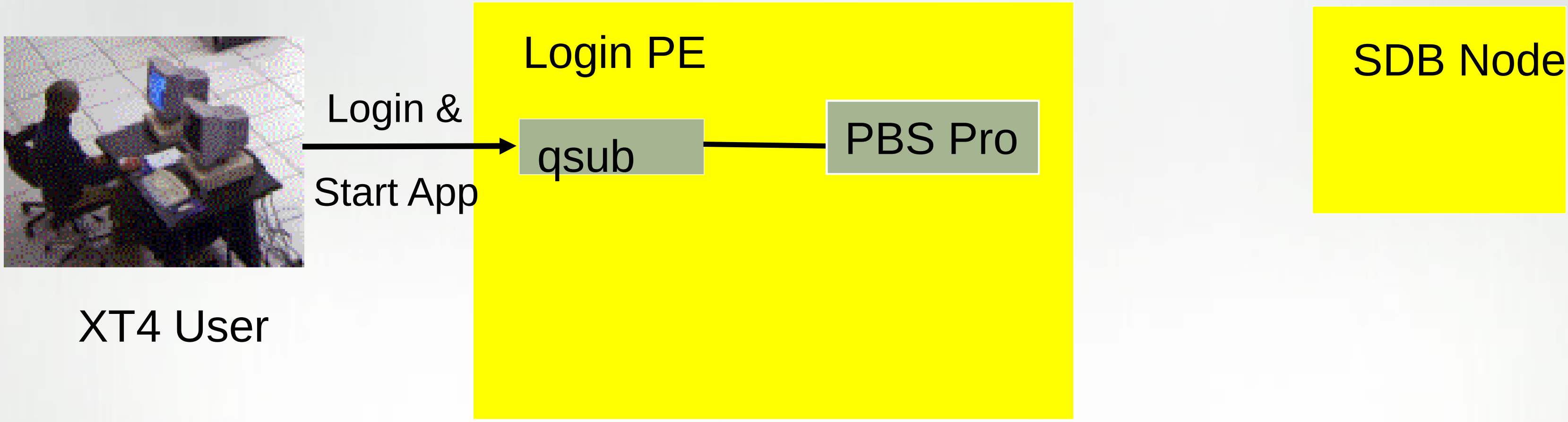


XT4 User

Login PE

SDB Node

Job Launch – The process



Login PE

qsub

PBS Pro

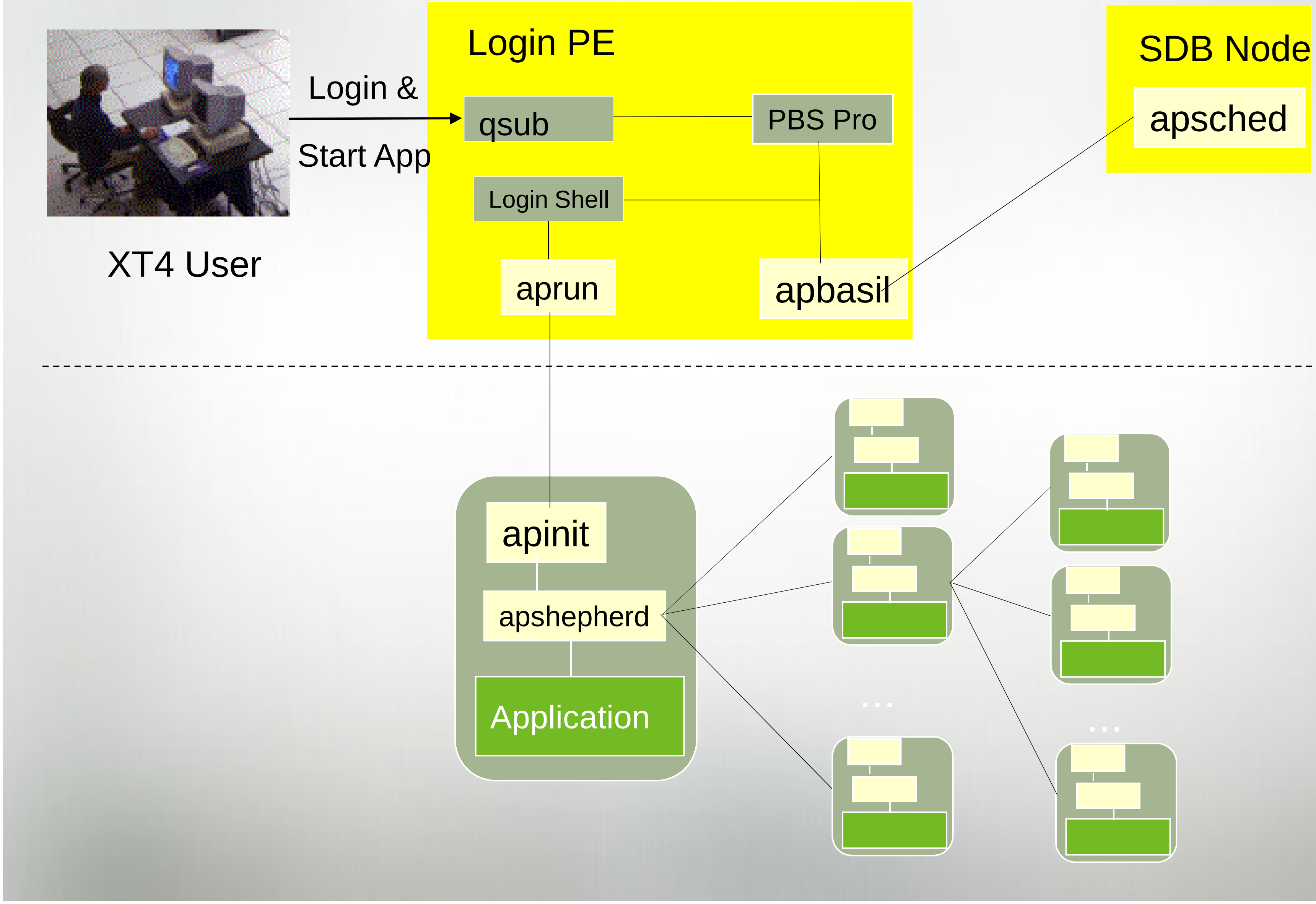
SDB Node

Login &

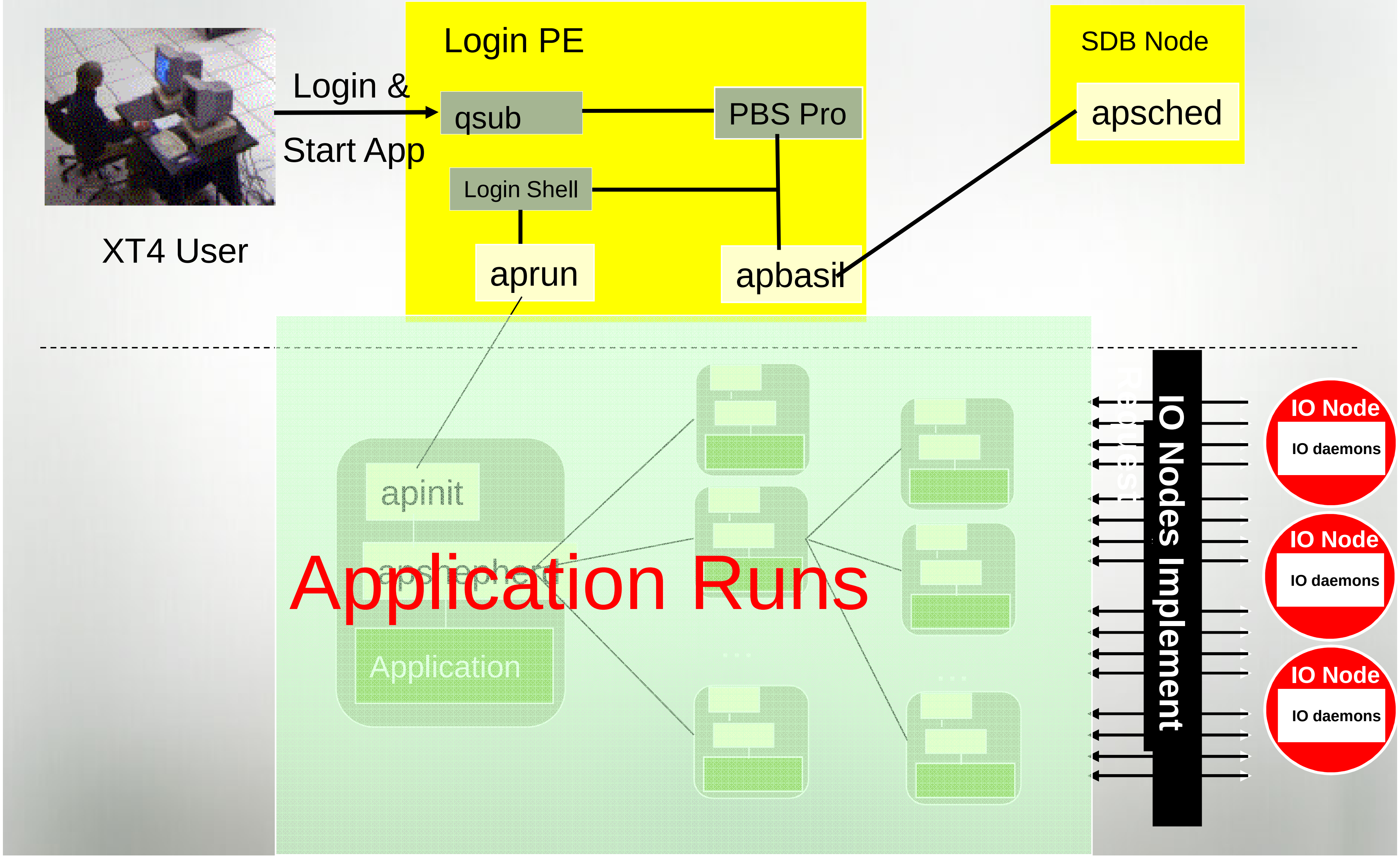
Start App

XT4 User

Job Launch – The process



Job Launch – The Process



XT4 User

Login &

Start App

Login PE

qsub

PBS Pro

Login Shell

aprun

apbasil

SDB Node

apsched

apinit

apshepherd

Application

Request

IO Nodes Implement

IO Node

IO daemons

IO Node

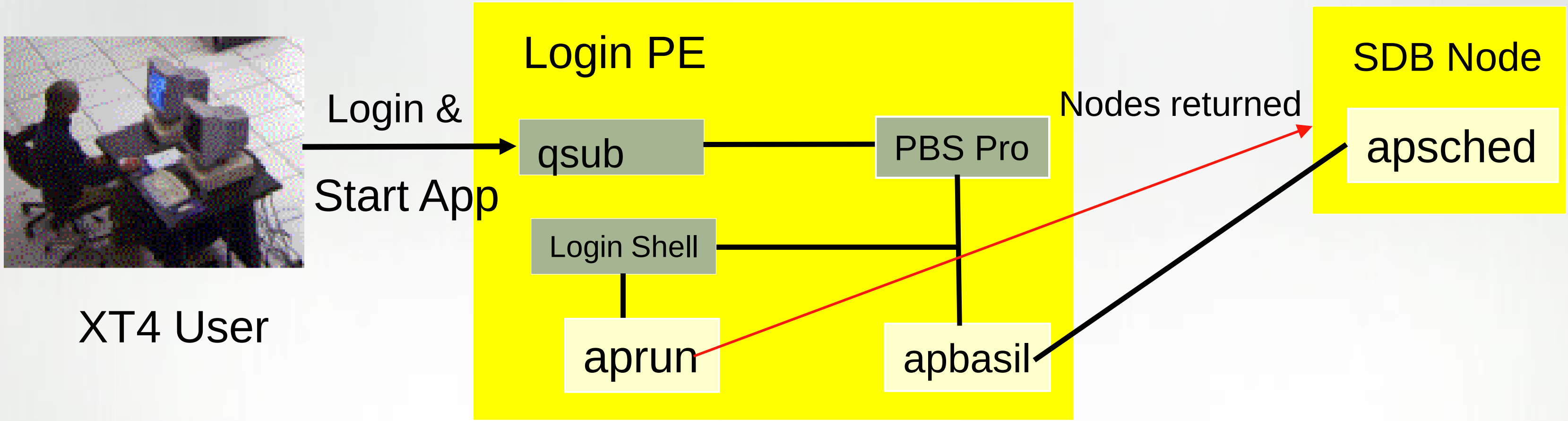
IO daemons

IO Node

IO daemons

Application Runs

Job Launch – The Process



Login &

Start App

XT4 User

Login PE

qsub

PBS Pro

Login Shell

aprun

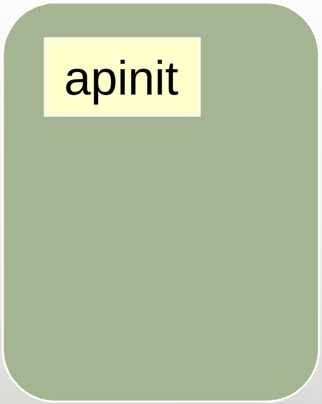
apbasil

Nodes returned

SDB Node

apsched

Job is
cleaned up



apinit

Creating Batch Jobs

- We use PBS to request resources
- These resources are then available for use to commands within the job (we must use aprun to access these resources).
  - Request CPU cores
  - Request time
  - Also for setting output files
- It is important that we make requests for the resources we require otherwise our parallel application will not launch efficiently.

Important PBS Flags

- -l mppwidth=X
  - Controls the number of nodes where alps will launch the parallel application (MPI)
- -l mppnppn=Y
  - Controls how many of these tasks are placed per node (MPI)
- -l mppdepth=Z
  - Controls how to spread out the tasks (required to request resource for OpenMP or threading)
- -l walltime=HH:MM:SS
  - How long you will need for the application
- -q NAME
  - Which queue to submit the job to

```
-l mppwidth=256
-l mppnppn=4
-l mppdepth=2
-q batch
```

Useful PBS Flags

- -o NAME
  - Where to place the standard output file
- -e NAME
  - Where to place the error file
- -j oe
  - Join the output and error file
- -A ACCOUNT
  - Which account to charge your job to if applicable
- -N NAME
  - A name for the run

Launching an Application

- We use ALPS to place tasks onto compute nodes
  - Application Level Placement Scheduler
- ALPS commands must be launched from a directory that is available to compute nodes
  - Does not need to contain the files required
  - /tmp or /scratch
- ALPS can only place tasks on the nodes reserved
  - You can use less
  - You cannot use more (Claim exceeds reservation's node-count)

Launching an Application

- -n X
  - Number of MPI (co-array or Shmem) ranks to place
- -N Y
  - Number of MPI ranks to place per node
- -d Z
  - Depth of the MPI rank (to produce spacing for memory or OpenMP)
  - For OpenMP still needs OMP_NUM_THREADS

- There are numerous indepth switches but the default is usually what you want).

```
aprun –n 256 –N 4 –d 2
aprun –n 512 –N 8 –d 1
```

Application Monitoring

- You can monitor the batch job by looking at qstat
  - qstat –a
  - qstat –f <JOBID>
  - qstat –u <USERNAME>
- You can monitor the application with apstat
- You can see where these are placed with xtnodestat

- If you have access to the node where aprun started you can look at the spooled job output (also available in home directory with the –k option to qsub).

xtnodestat

```
Current Allocation Status at Mon Sep 21 04:03:58 2009
     C0-0
  n3 --------
  n2 --------
  n1 -----a--
c2n0 --------
  n3 SS------
  n2   ------
  n1   ------
c1n0 SS------
  n3 SSSA----
  n2    A----
  n1    A----
c0n0 SSSA----
    s01234567
Legend:
   nonexistent node                 S  service node
;  free interactive compute CNL     -  free batch compute node CNL
A  allocated, but idle compute node ?  suspect compute node
X  down compute node                Y  down or admindown service node
Z  admindown compute node           R  node is routing
Available compute nodes:          0 interactive,         71 batch
Job ID     User       Size   Age              command line
--- ------ --------   -----  ---------------  ---------------------------------
a   4734294 freddy     1      0h00m            xsort
```

THE SUPERCOMPUTER COMPANY

Running on XT Compute Nodes

Questions / Comments

Thank You!