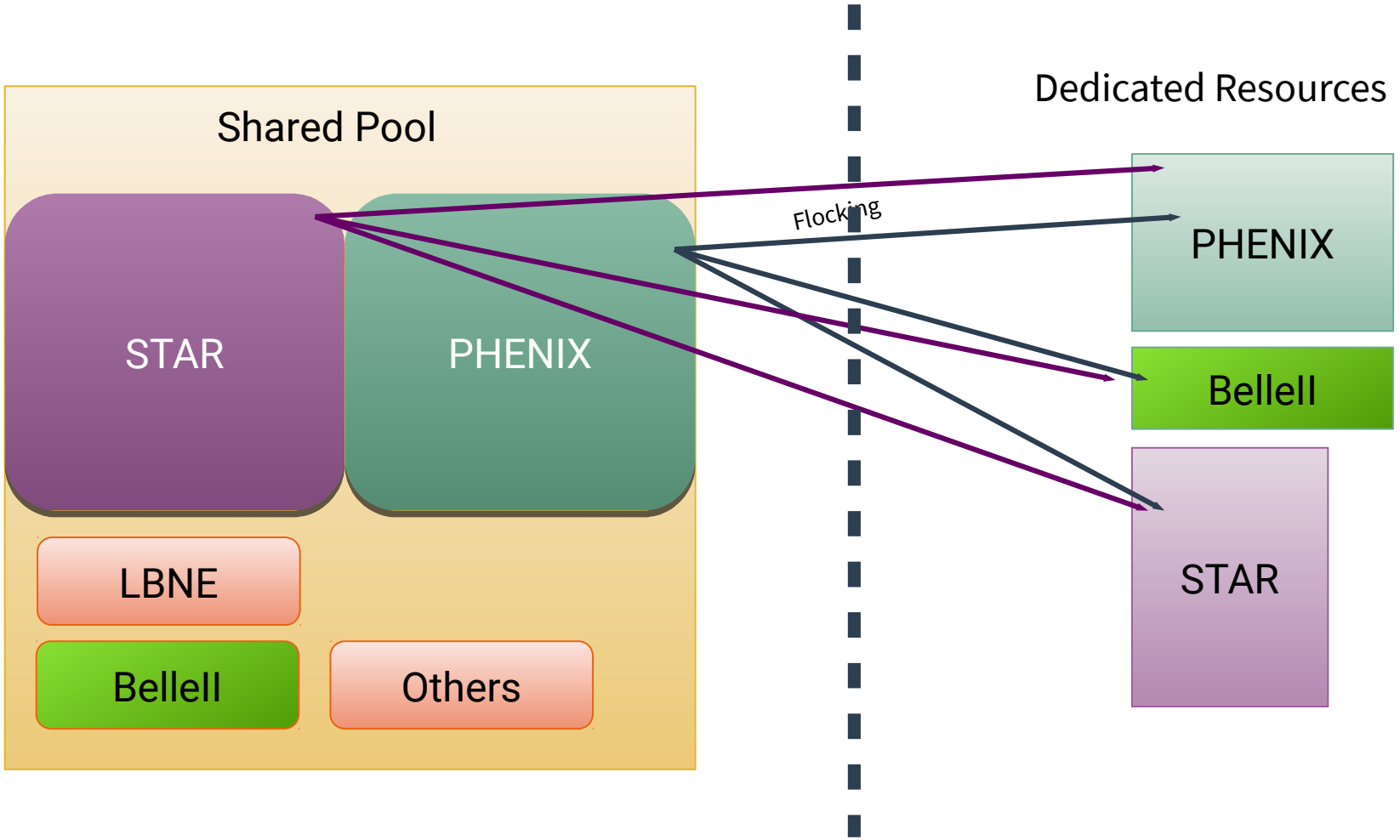
BNL Shared Analysis Pool

Model


Shared Pool
STAR
PHENIX
LBNE
BelleII
Others
Flocking
Dedicated Resources
PHENIX
BelleII
STAR

Accounting Groups

- **Virtual allocations of shared pool**
  - Segregate experiments and queues into groups
    - group_<experiment>.<queue>
  - Spillover between groups enabled
    - Full queue steady-state will reflect the allocations
    - Any group without jobs will have their resources be shared proportionally by the others
  - Fair-share done within groups (usual priority decay)
  - Next slot goes to most deprived user in most deprived group group

Dedicated Resources

- **Pools per experiment**
  - Can retain resources for special use cases that don't fit policy / hardware of shared pool
  - Pools will be back filled with jobs from the shared pool
    - Each experiment's interactive nodes will flock *first* to their own dedicated resources then to others
  - Need to keep "reasonably" full to justify

Job Submission

- **User interactive nodes move into shared pool**
  - Remember: can flock back to dedicated resources
- **Special interactive (CRS / phnxsub / anything else…) exist in dedicated pool**
  - Limited users, special uses, custom policies

User-Facing Consequences

- **CPU_Experiment / CPU_Type go away**
  - Requirements / Rank expressions need to change accordingly
  - Group assignment done at the schedd-level; users don't need to manually add
- **Changes to policy r.e. runtimes / limits**
  - As agreed
  - Any special group/subgroups can be added
    - e.g. +LongJob=true

Monitoring

- **Will determine—with your input—salient data to monitor**
  - We want to be able to ensure that user experience remains good
  - Will try to monitor utilization, latency, efficiency, etc…

Carrots

- **Shared pool will grow with resources beyond traditional HTC**
  - Access via condor job-router to other (HPC) resources in the plans
- **Quicker access to more resources**
  - Surplus-sharing is faster and more efficient than flocking
- **Shared-facility model aligns with DOE priorities**
- **As tenants→N, efficiency goes up**

Sticks

- **Multi-tenant facility requires agreement on policy parameters**
- **Slightly stricter resource-control**
  - Will monitor, but will likely need cgroups for per-job iops, disk space, etc…
- **Requires more careful planning of special user-cases / moving these into smaller dedicated-resources**

Open Questions

- **Points to discuss**
  - Preemption vs. Attrition
  - Time limit (3 day should cover most analysis use, but negotiable)
    - Per-queue a possibility
  - Resource limits (high memory nodes?)
  - What data to monitor