U.S. DEPARTMENT OF ENERGY | Office of Science

ATLAS EXPERIMENT

Streaming in ATLAS

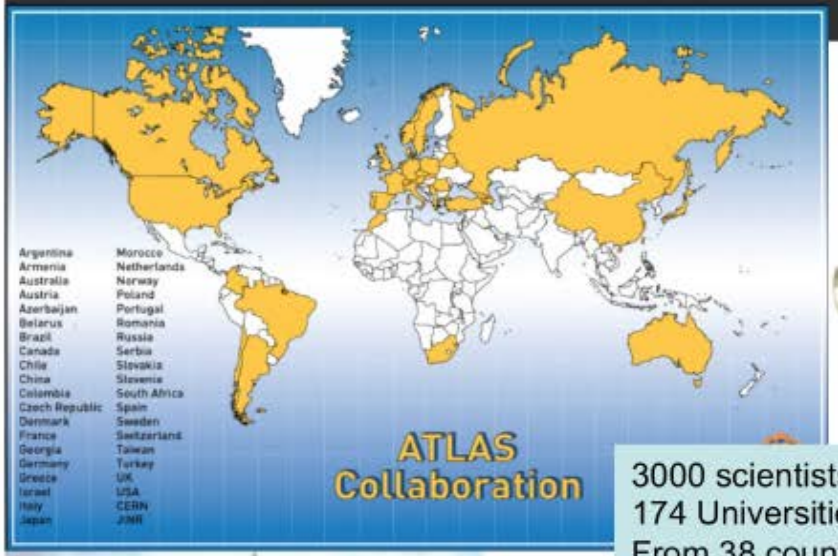
Vakho Tsulaia (LBNL), Torre Wenaus (BNL)

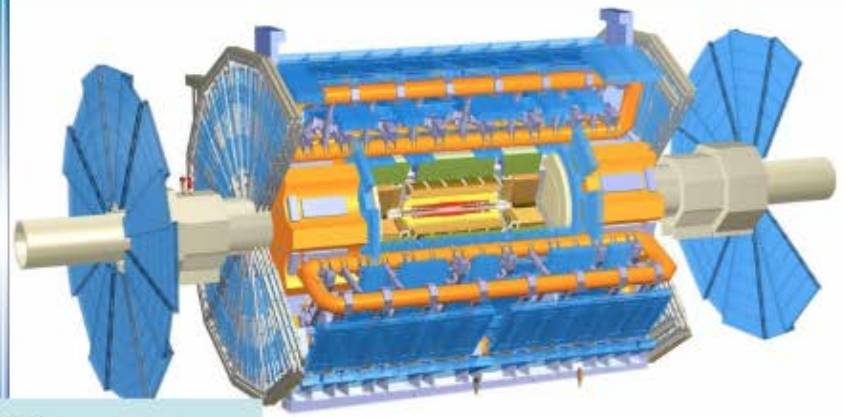
STREAM 2016
Tysons, VA
March 22, 2016

BROOKHAVEN NATIONAL LABORATORY

BERKELEY LAB
Lawrence Berkeley National Laboratory

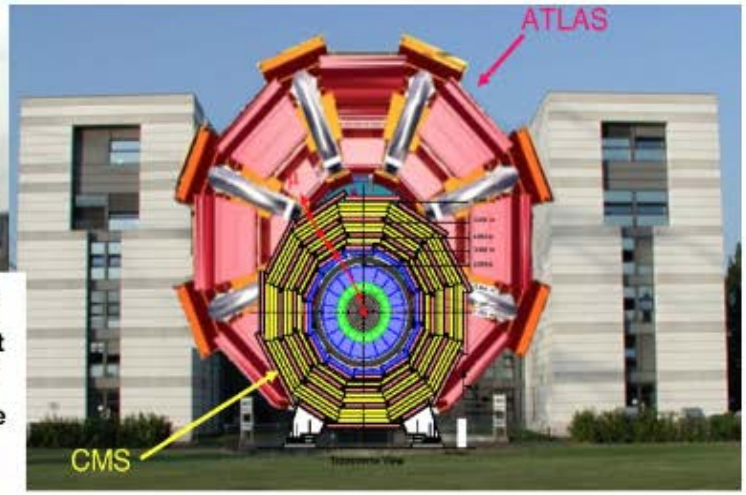
The ATLAS Experiment at the LHC





3000 scientists
174 Universities and Labs
From 38 countries
More than 1200 students

- **ATLAS has 44 meters long and 25 meters in diameter, weighs about 7,000 tons. It is about half as big as the Notre Dame Cathedral in Paris and weighs the same as the Eiffel Tower or a hundred 747 jets**



The Nobel Prize in Physics 2013
François Englert, Peter Higgs

The Nobel Prize in Physics 2013

François Englert | Peter W. Higgs

The Nobel Prize in Physics 2013 was awarded jointly to François Englert and Peter W. Higgs *"for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider"*

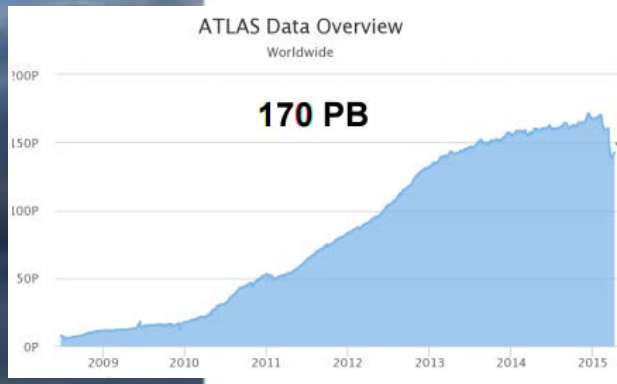
Large Scale Data Intensive Processing: The LHC Data Torrent

New physics rate ~ 0.00001 Hz

Event Selection :

1 in 10,000,000,000,000

Like looking for a single drop of water from the Geneve Jet d'Eau over 2+ days



ATLAS: ~1PB raw data/s off the detector filtered to 1-2 GB/s recorded

ATLAS Computing Essentials

- Globally distributed by necessity: computing follows the people and support dollars
  - The ATLAS Grid would be about #27 on the HPC Top 500
  - And it isn't enough: big push into opportunistic resources
- 140+ heterogeneous resources sharing 170PB and processing exabytes per year, with a few FTEs of operations effort
- Our ability to do that is grounded in:
  - **Excellent networking**, the bedrock enabler for the success of LHC computing since its inception
  - **Workflow management** that is intelligent, flexible, adaptive and intimately tied to **dataflow management**
  - Dataflow management must minimize storage demands by **replicating minimally and intelligently**, using our **networks to the fullest** by sending **only the data we need, only where we need it**

From fine grained steering to fine grained dataflow

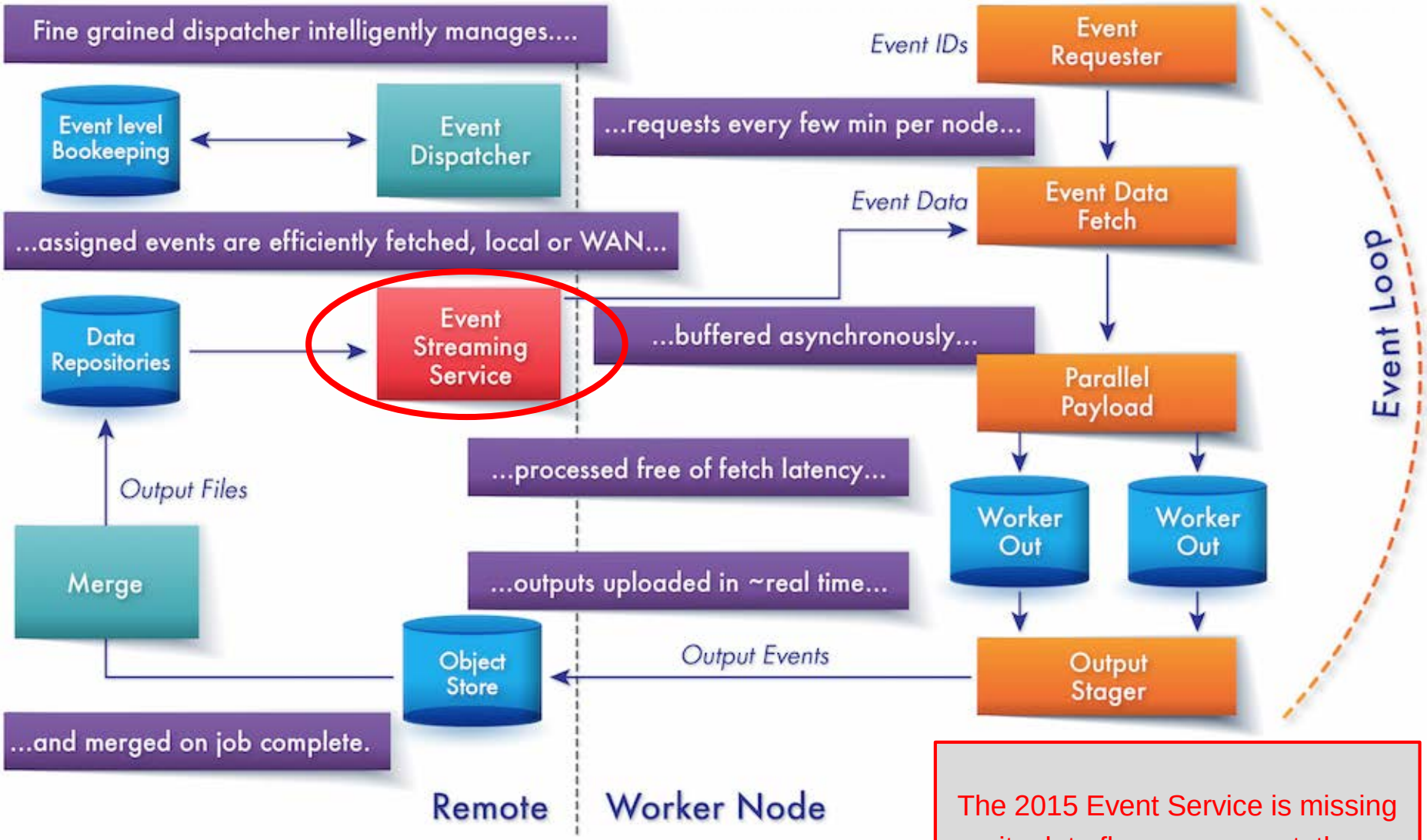
- The ATLAS **Event Service (ES)**: a new approach to HEP processing
  - **Quasi-continuous event streaming** through worker nodes
  - Agile, dynamic tailoring of workloads to fit the scheduling opportunities of the moment (**HPC backfill**)
  - Loss-less termination (**EC2 spot market node disappearance**)

- Exploit event processors fully and efficiently through their lifetime
  - **Real-time delivery of fine-grained workloads to running application**

- Decouple processing from chunkiness of files, from data locality considerations and from WAN latency

- Stream outputs away quickly
  - **Minimal local storage demands**

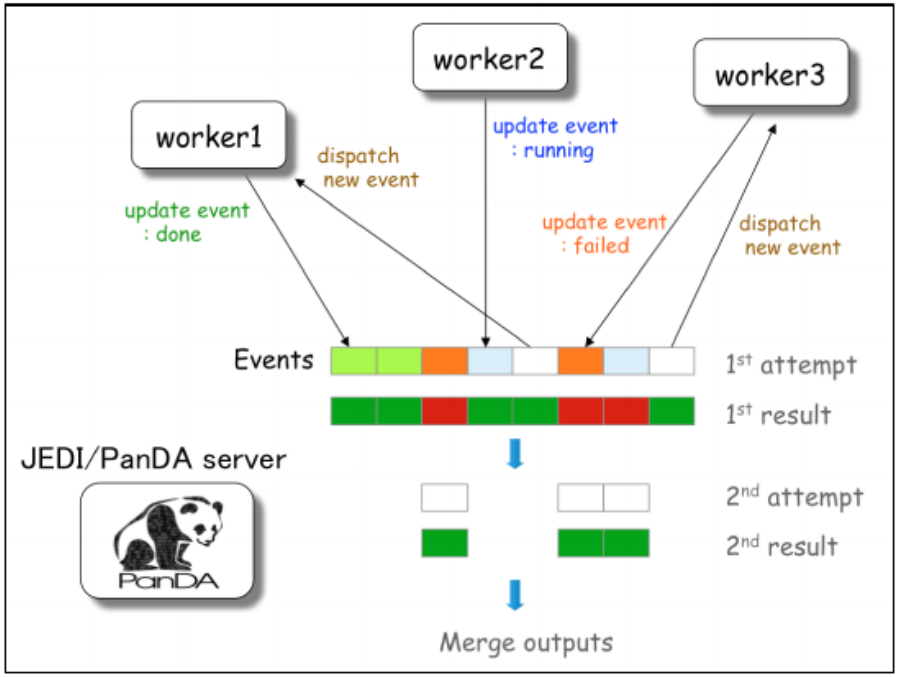
Event Service in 2015



The 2015 Event Service is missing its data flow component, the Event Streaming Service

ES Building Blocks

- **The ES Engine**: PanDA Distributed Workload Manager
  - **JEDI** extension to PanDA adds flexible task management and fine-grained dynamic job management
- **Parallel payload**
  - Efficient usage of CPU and memory resources on the compute node
  - Whole-node scheduling
- **Remote I/O**
  - Efficient delivery of event data to compute nodes
- **Object Stores**
  - Efficient management of outputs produced by the ES



Yoda: Event Service on Supercomputers

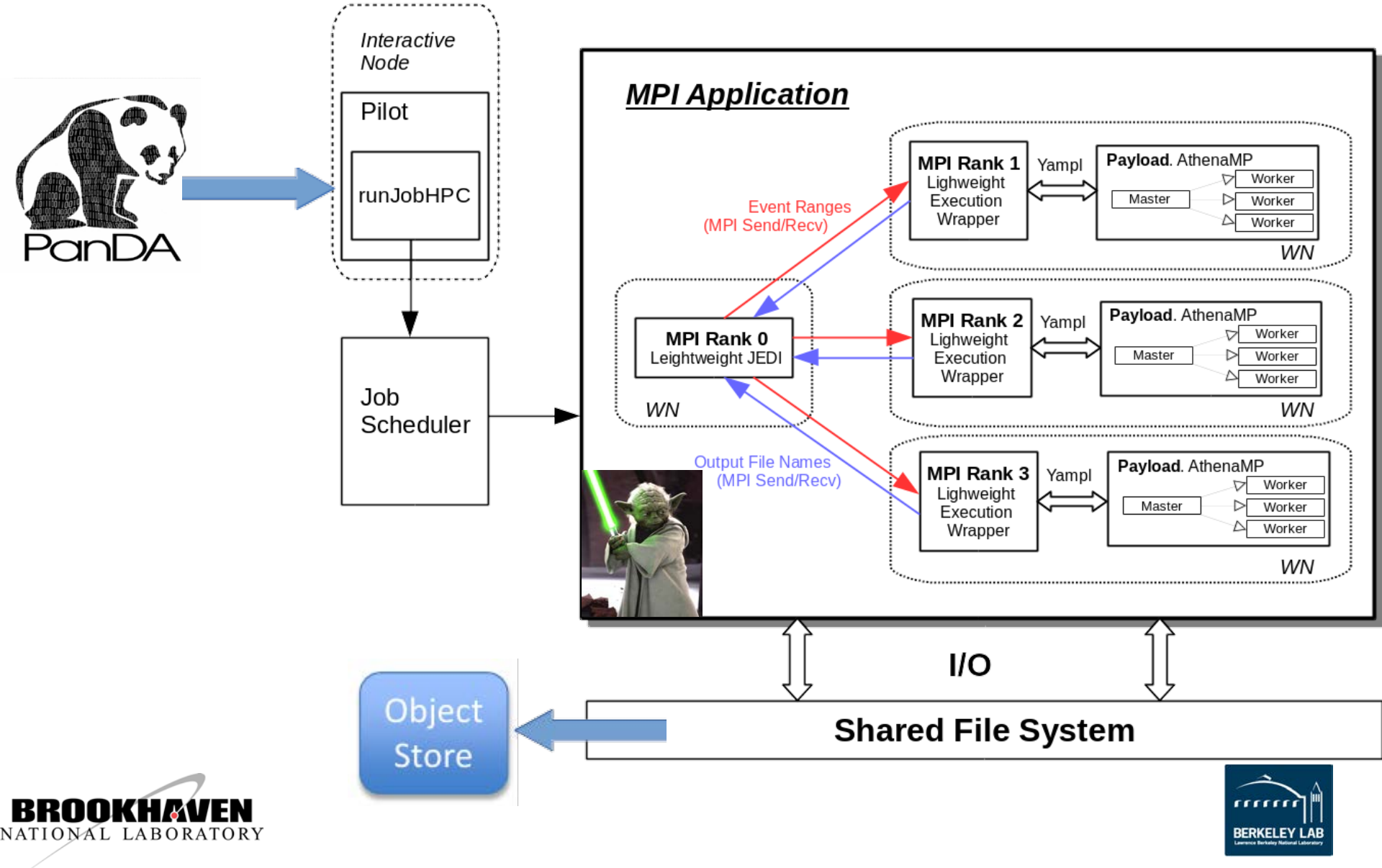
- While PanDA was originally developed for the Grid, BigPanDA and ATLAS have extended it to operate also as an HPC internal system
  - Designed for efficient and flexible resource allocation and management of **MPI-based parallel workloads within HPC**

- **Yoda** - HPC-internal version of PanDA - leverages the experience acquired in massively scaled data Intensive processing for efficient utilization of a single massively scaled HPC machine
- The PanDA team is working with computing specialists at **NERSC**, **OLCF** and **ALCF** on implementing several approaches towards fine-grained, adaptive, flexible workflows to achieve the highest possible system utilizations
  - Both **backfill** and **scheduled** allocation modes

BROOKHAVEN NATIONAL LABORATORY

BERKELEY LAB
Lawrence Berkeley National Laboratory

Yoda. Schematic view

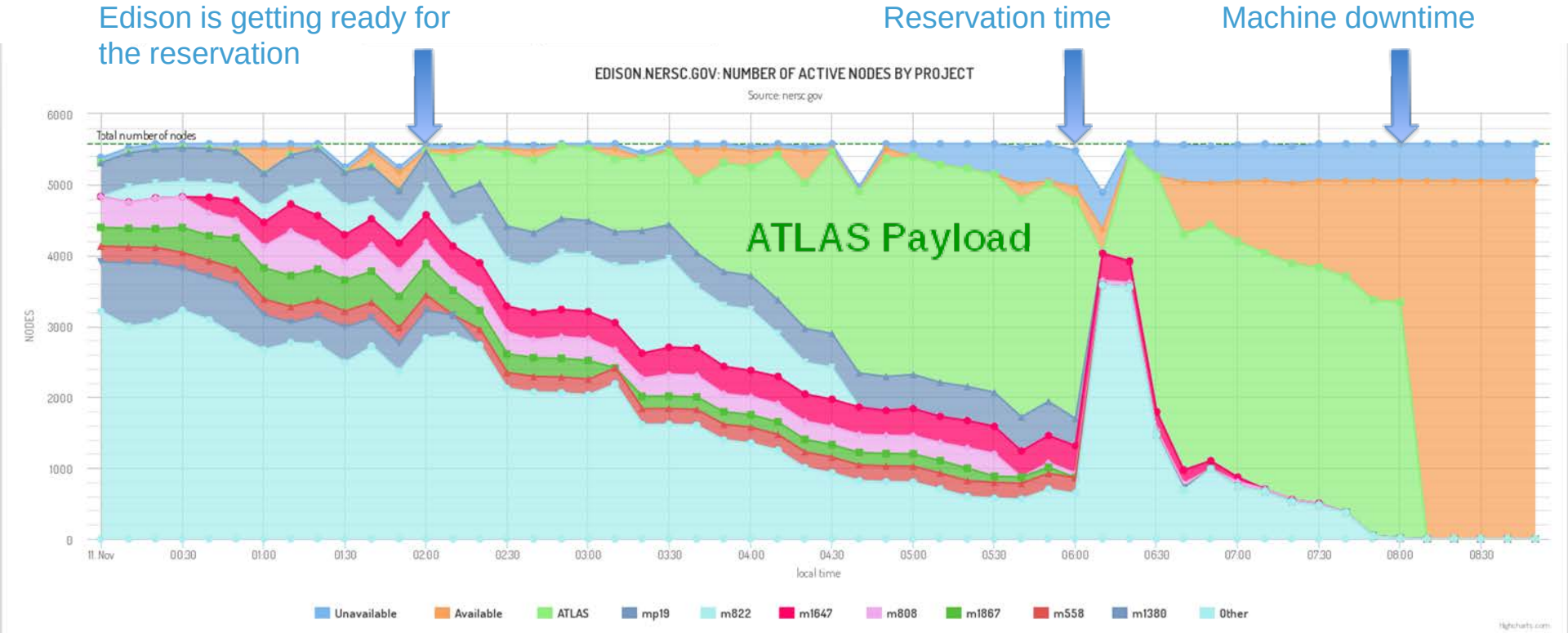


BROOKHAVEN NATIONAL LABORATORY

BERKELEY LAB

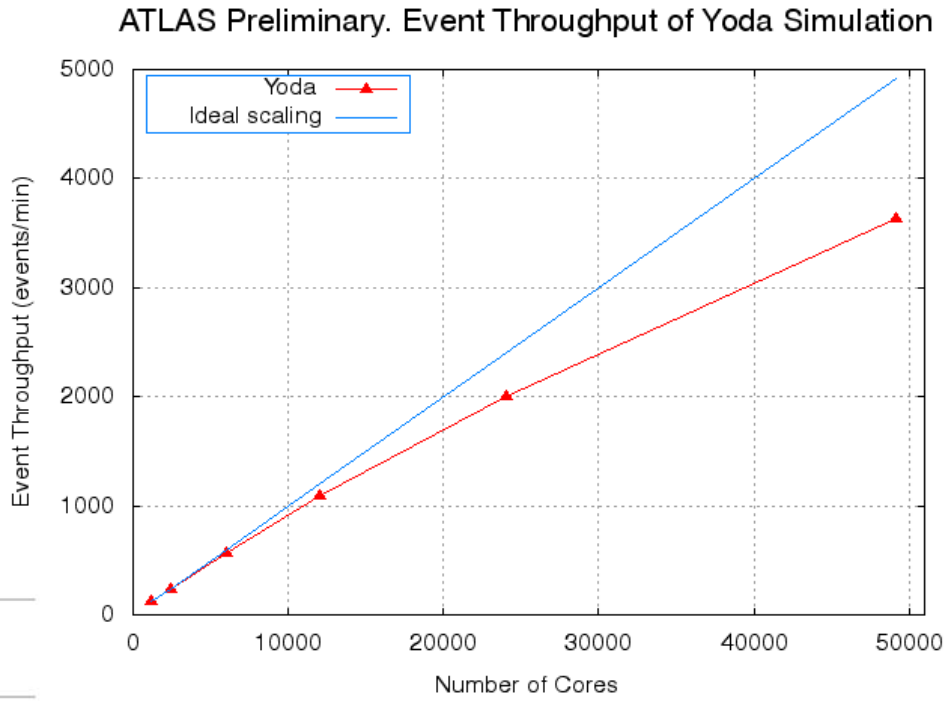
Yoda scavenging resources

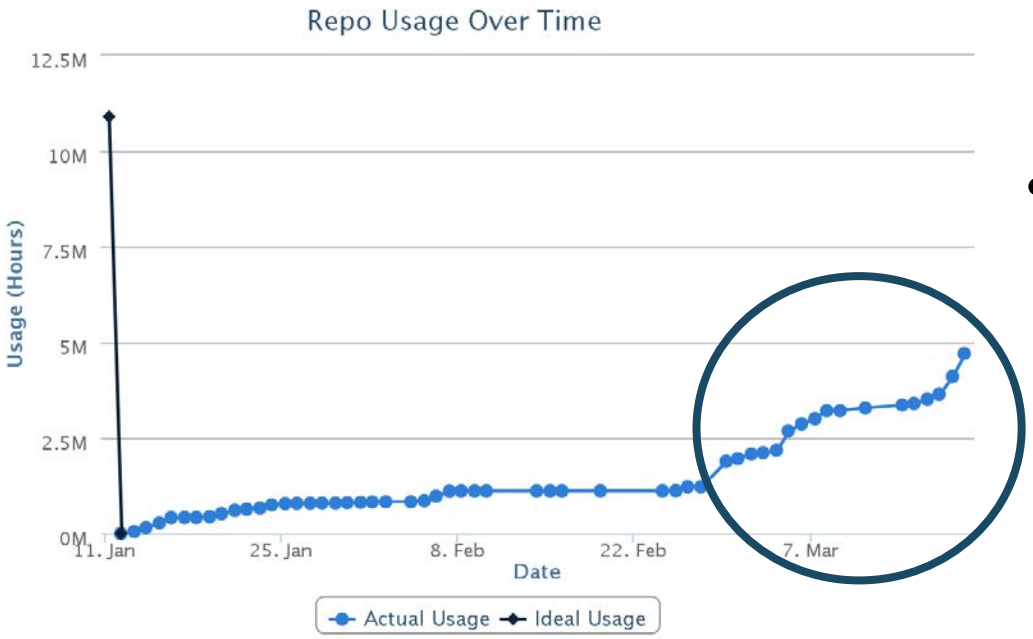
- **“Killable queue test” on Edison HPC, 2014**
- As the machine is emptied either for downtime or for large “reservation”, the killable queue makes transient cycles available
- Yoda uses the resources until the moment they vanish, and refills them when they appear again



Yoda running at scale

- Geant4 Simulation of the ATLAS Detector on Edison HPC
- **Yoda running with 50K parallel processes simulated 220K full ATLAS events in 1hr**





- **Yoda running ATLAS Simulation workloads in production consumed 3.5M CPU-hours in March 2016**

BERKELEY LAB

From ES to Event Streaming Service (ESS)

- The Event Service can integrate perfectly with a similarly event-level data delivery service, the ESS, that responds to requests for "science data objects" by intelligently marshaling and sending the data needed

- Such service can encompass
  - CDN-like optimization of data sourcing "close" to the client
  - Knowledge of the data itself sufficient to intelligently skim/slim during marshaling
  - Servicing the request via processing on demand rather than serving pre-existing data

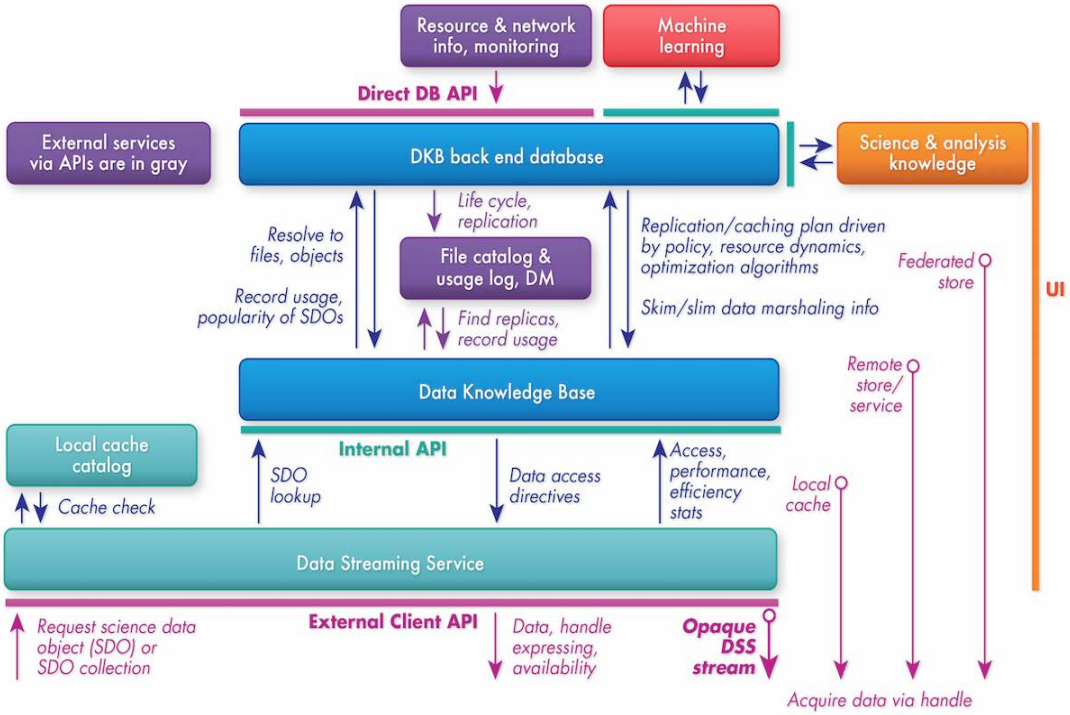
- We have to build it as an exascale system

Building the ESS

Two primary components

- **Data Streaming Service**
  - CDN-like intelligence in Finding the most efficient Path to data
  - Minimal replication
  - Data marshaling
  - Smart local caching

- **Data Knowledge Base**
  - Dynamic resource landscape
  - Science data object knowledge
  - Analysis processes and priorities



BROOKHAVEN NATIONAL LABORATORY

BERKELEY LAB

Conclusion

- ATLAS pushes today the bounds of data intensive science with exascale processing workflows on a 170PB data sample across >100 global sites
- ATLAS is moving to new, fine grained processing model to sustain the growth of its science and its computing needs
- The Event Service, built and commissioned, is now running ATLAS production workloads at large scale
- The Event Streaming Service is currently at the design/prototyping stage
  - **Looking for tools to build ESS that streams our Exabyte-scale data flows through the ES!**