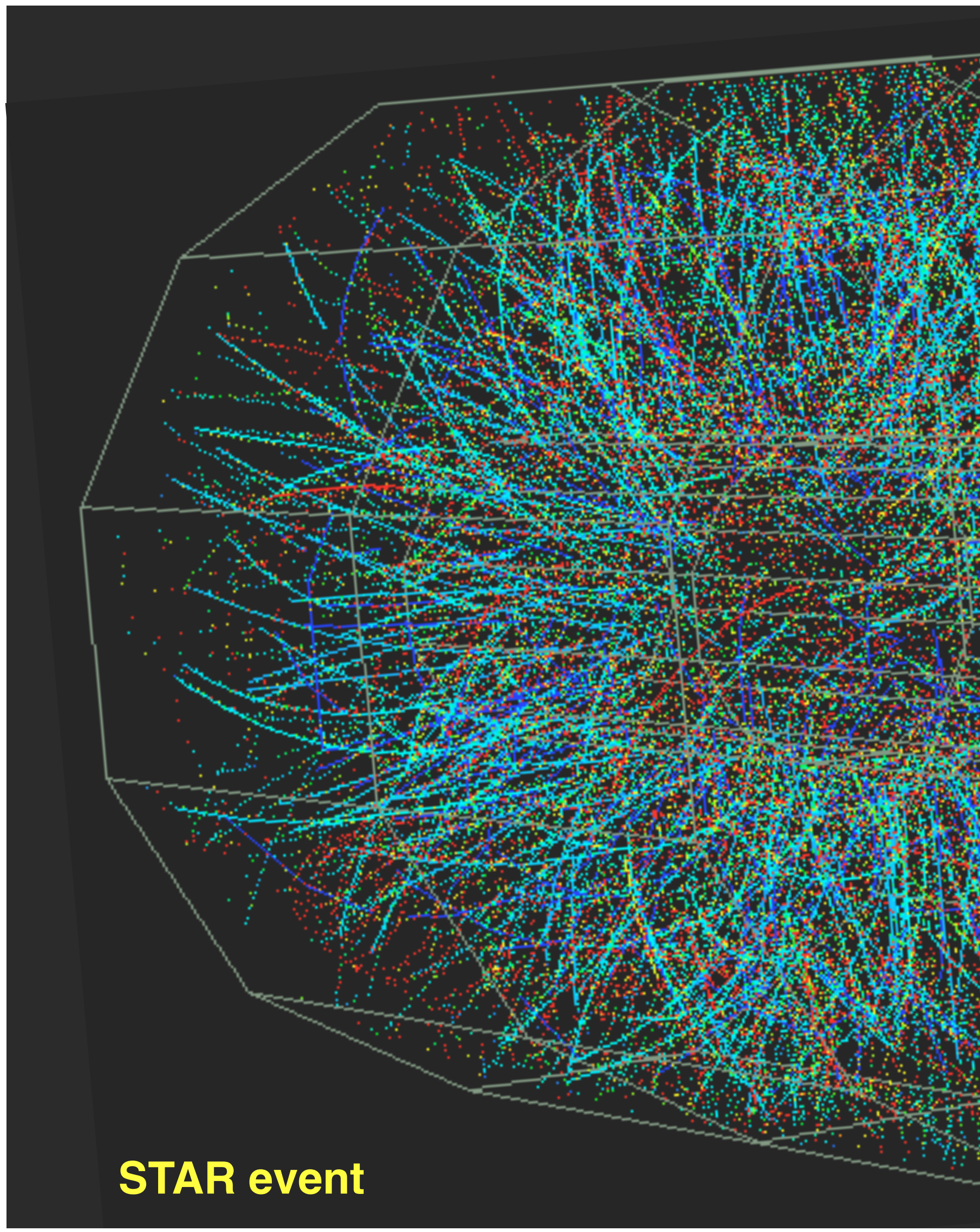
RHIC Real time data reconstruction using Magellan cloud computing



Jan Balewski
for STAR Collaboration

2011 OSG All Hands

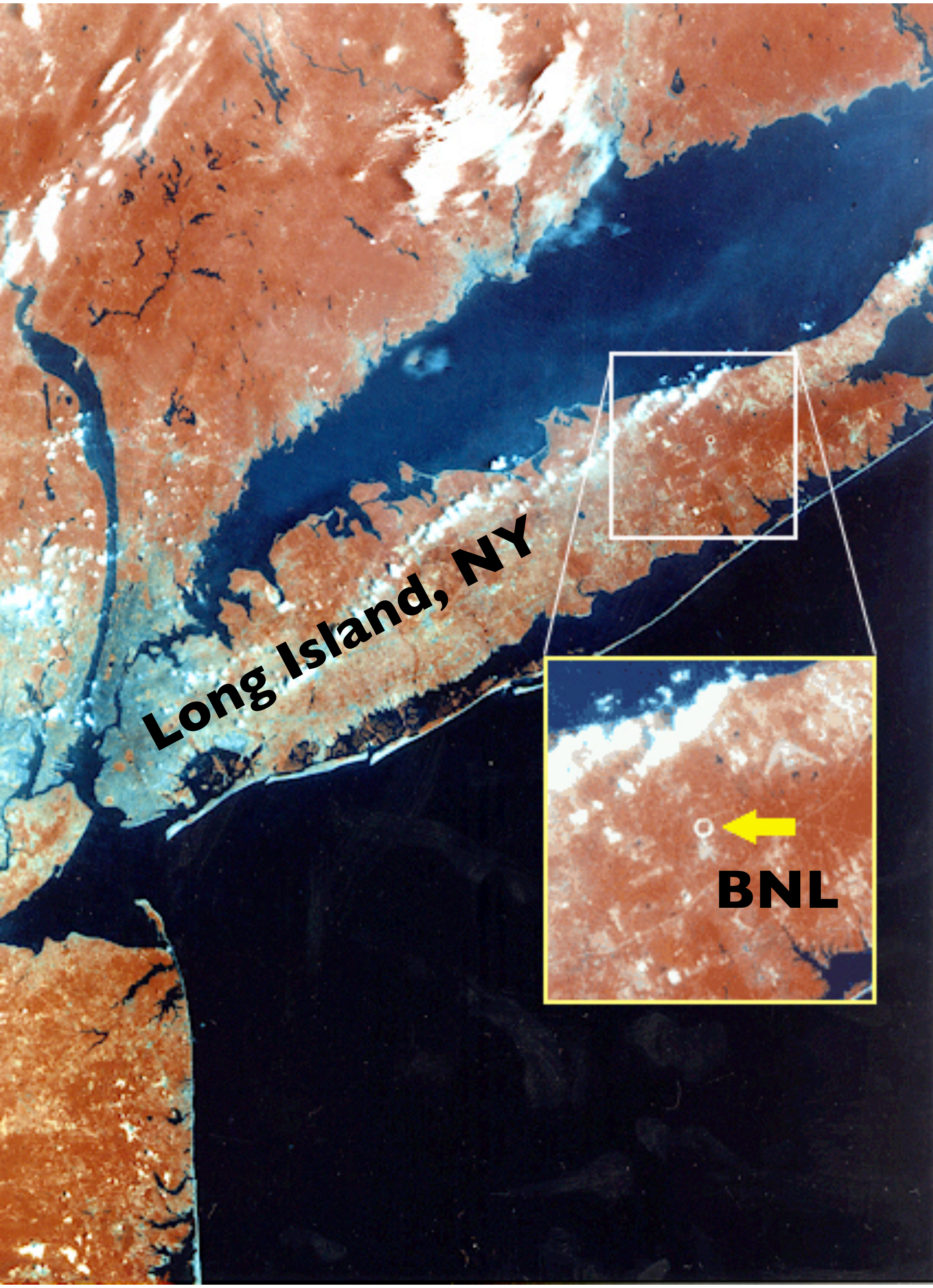
March 7-11, 2011

Harvard Medical School, Boston

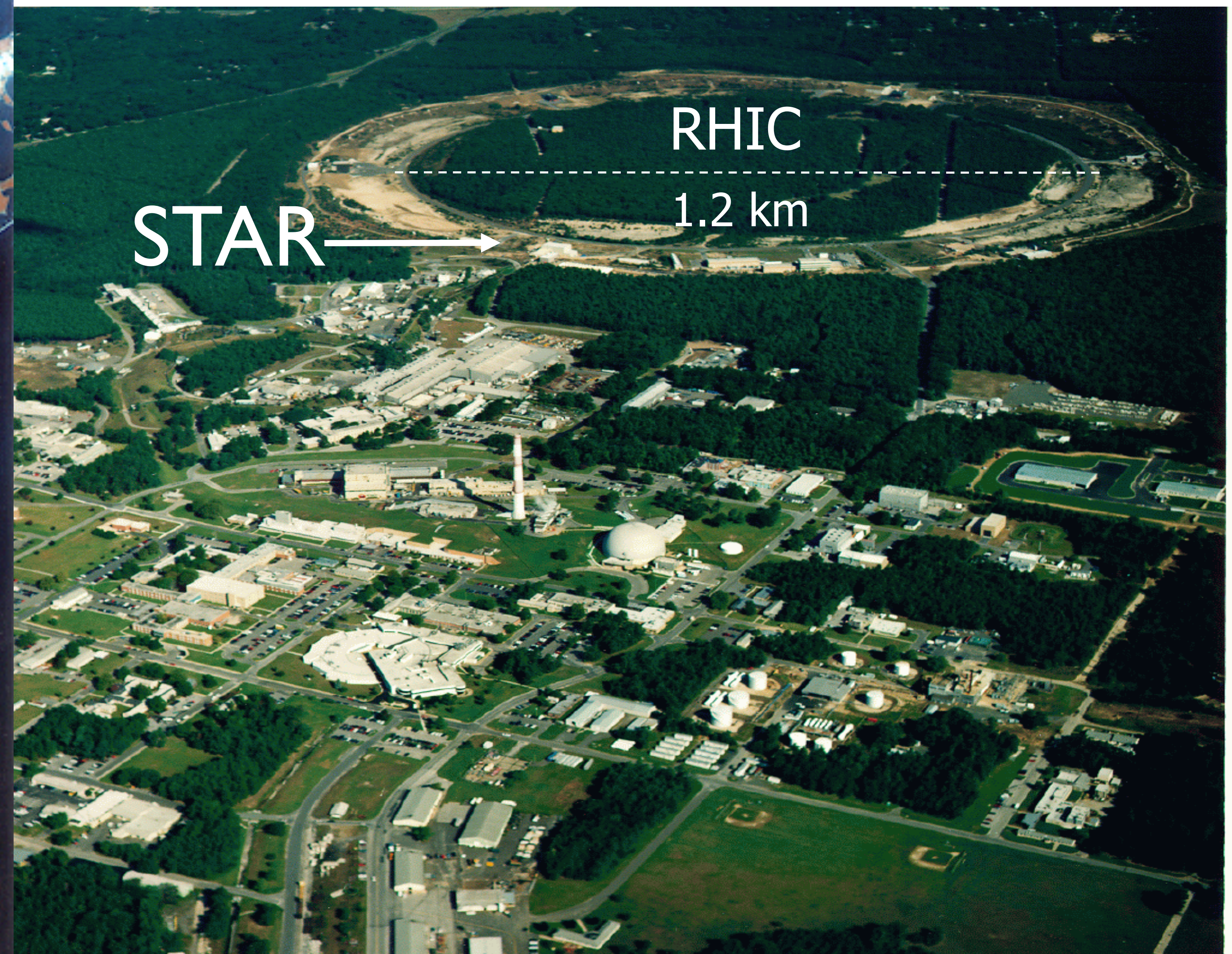
Outline

- STAR experiment at RHIC
- Computing requirements for real data analysis
- STAR encounters with Cloud-like computing
- Deployment of real time data processing
- Benefits of "instantaneous" data analysis
- Summary + ...

STAR experiment at RHIC

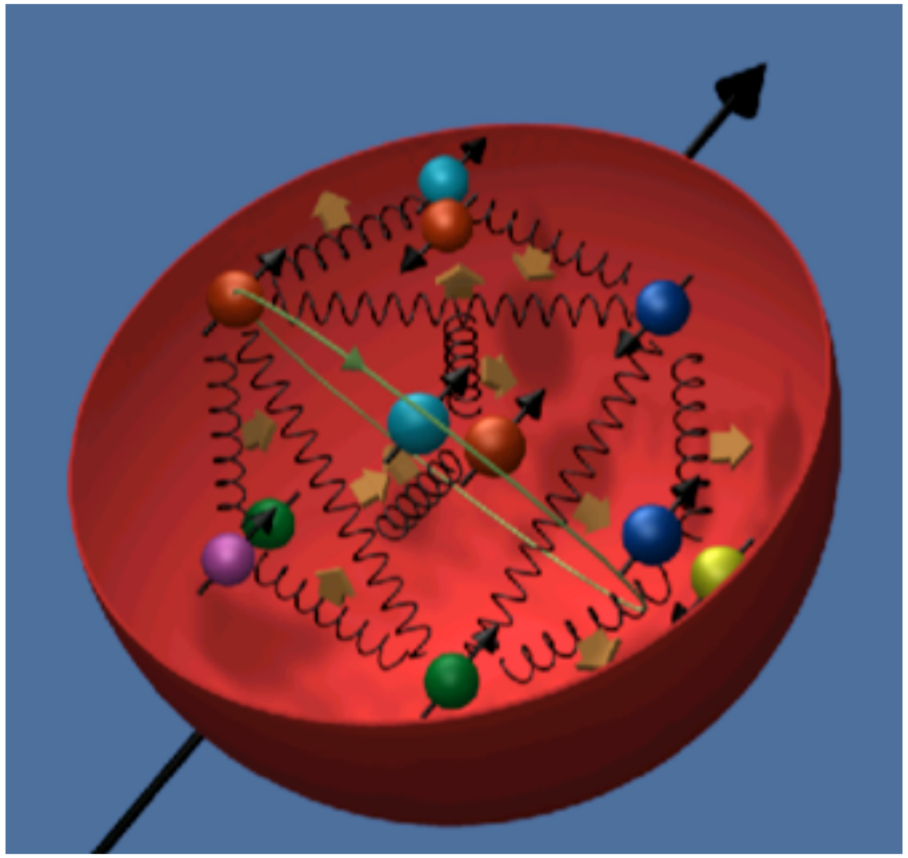


~600 collaborators from
~50 institutions and ~12 countries

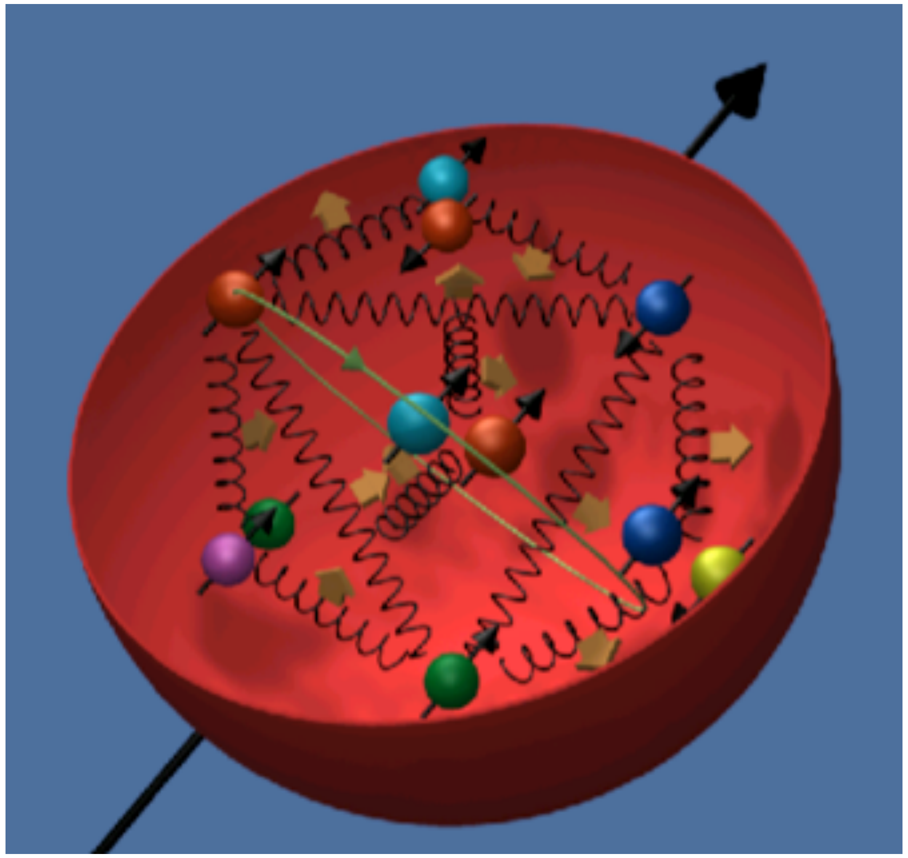


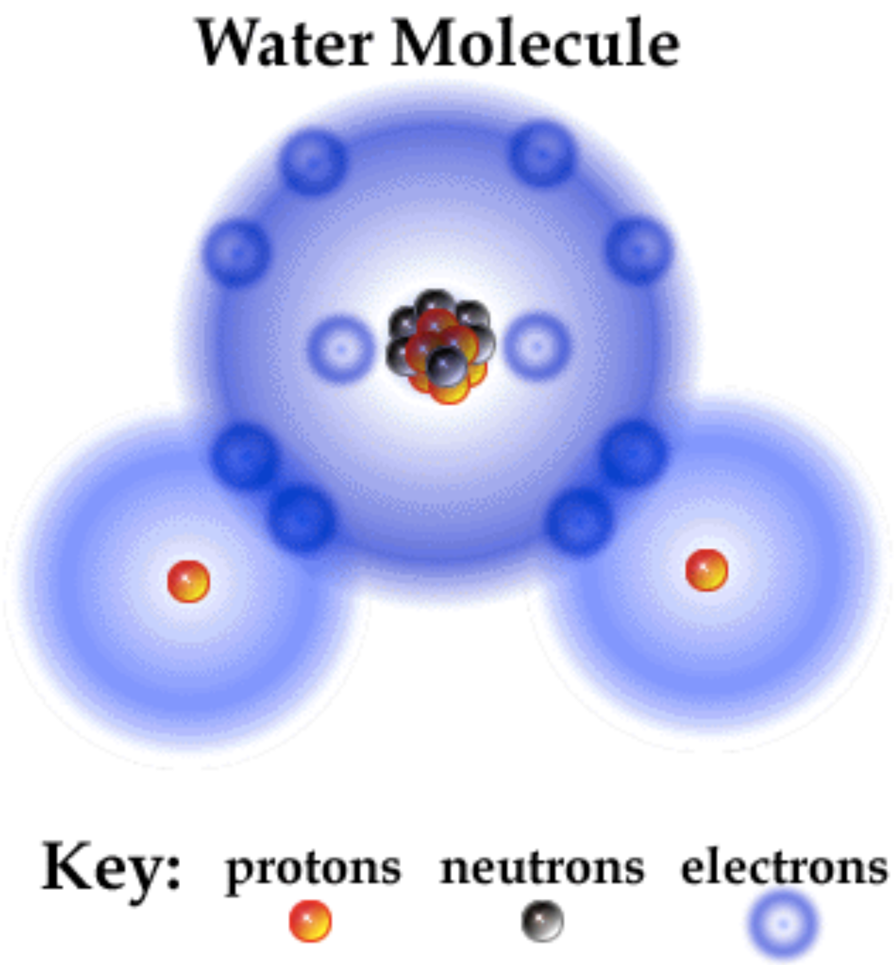
Brookhaven National Laboratory, Upton NY, USA

Explore properties of proton spin

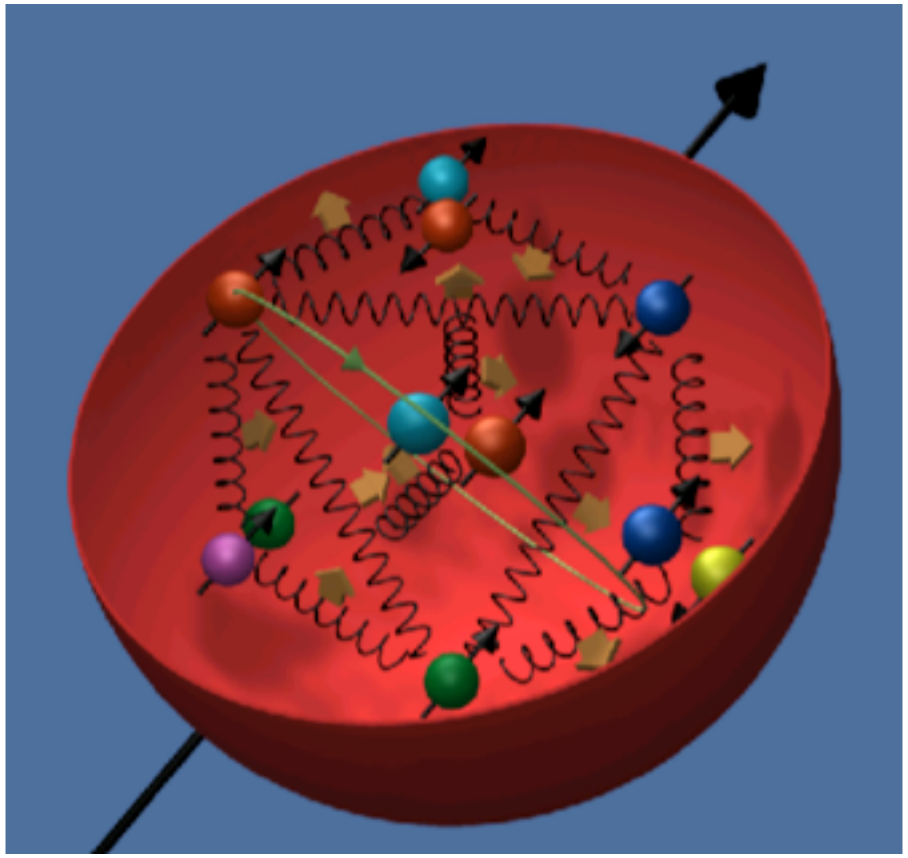


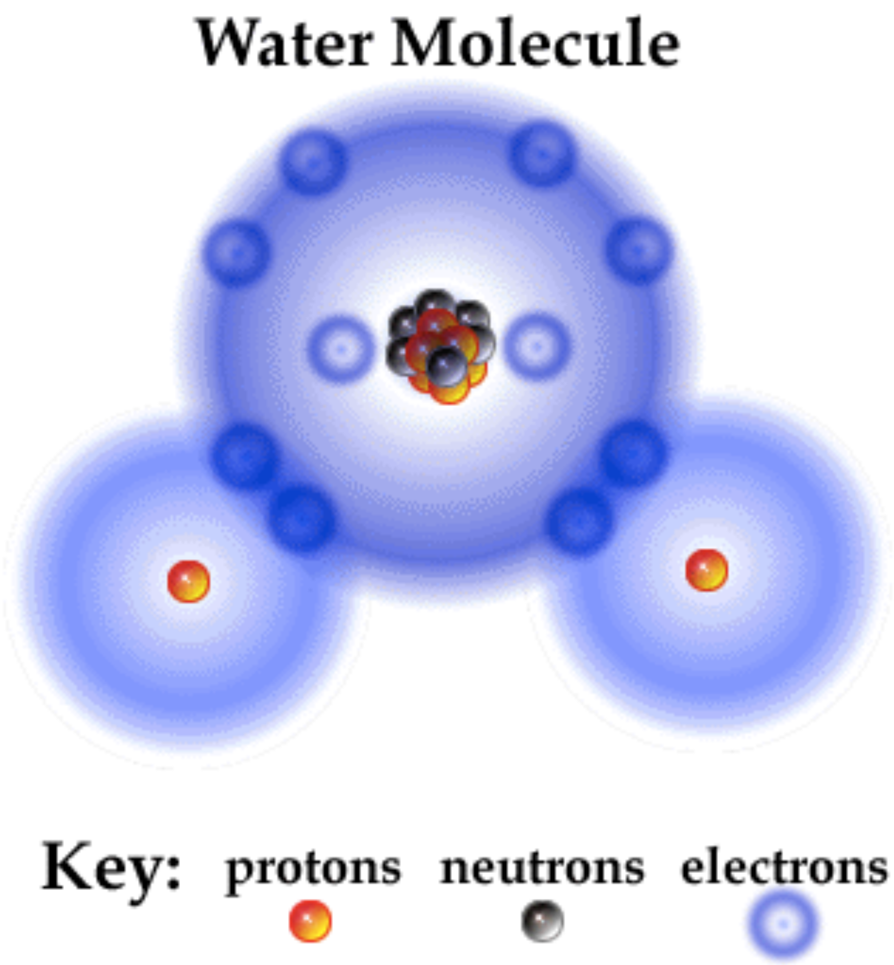
Explore properties of proton spin



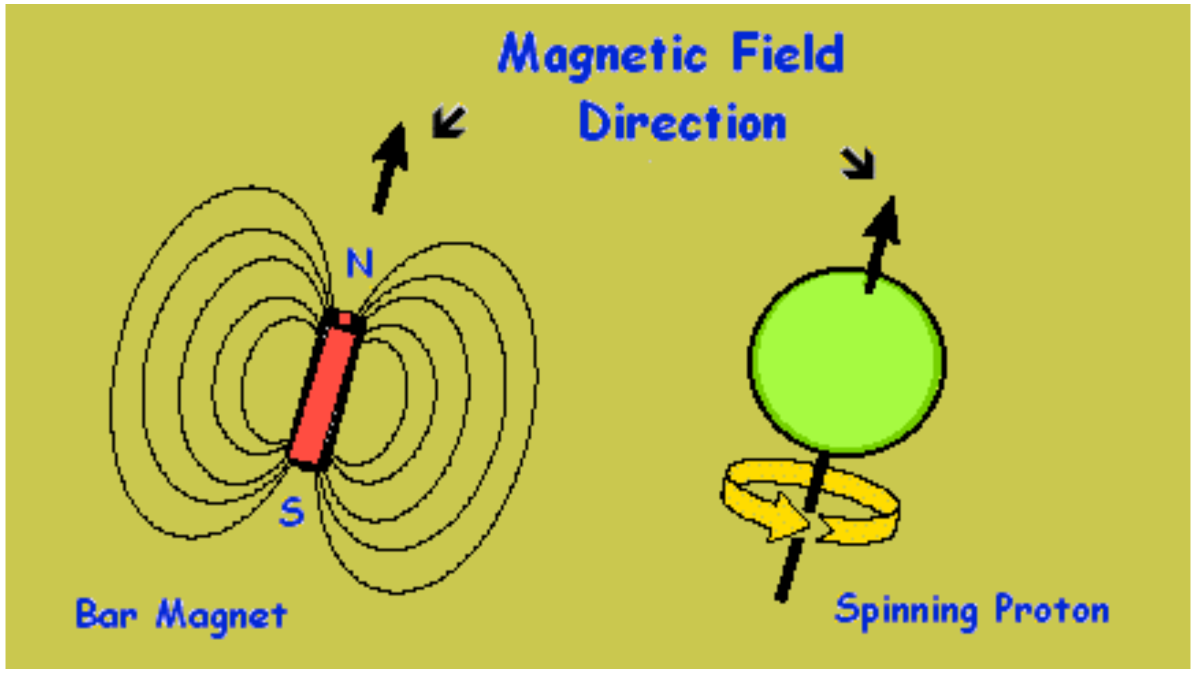


Explore properties of proton spin





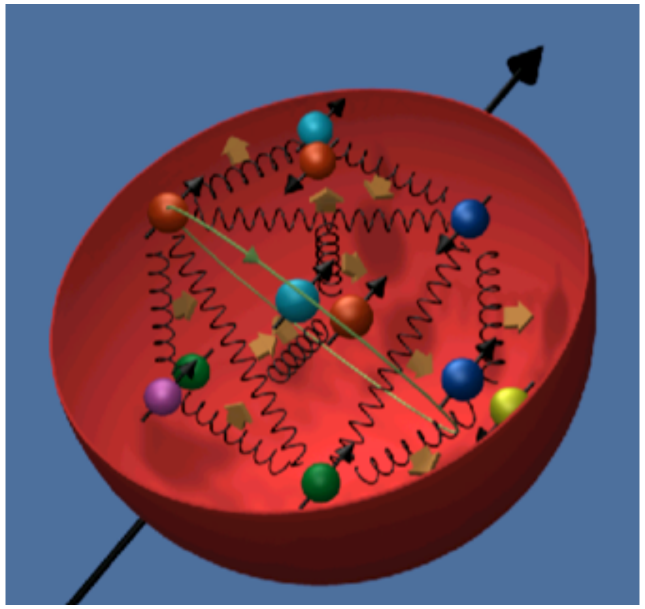
Magnetic Resonance Imaging



Explore properties of proton spin

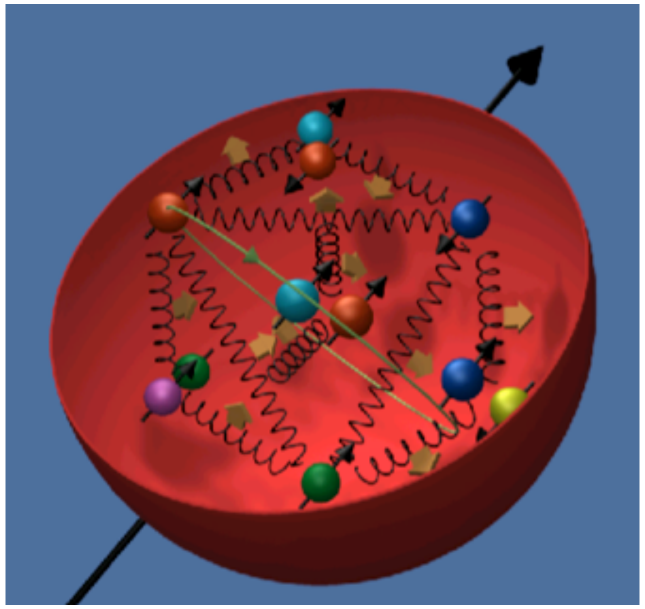
"Exploring the mystery of **proton spin** has been one of the key scientific research goals at RHIC," said **Steven Vigdor, Brookhaven's Associate Laboratory Director** for Nuclear and Particle Physics. .... **The W boson measurements [will help us] ... in quantitative understanding of proton spin structure and dynamics.**"

http://www.bnl.gov/bnlweb/pubaf/pr/PR_display.asp?prID=1232

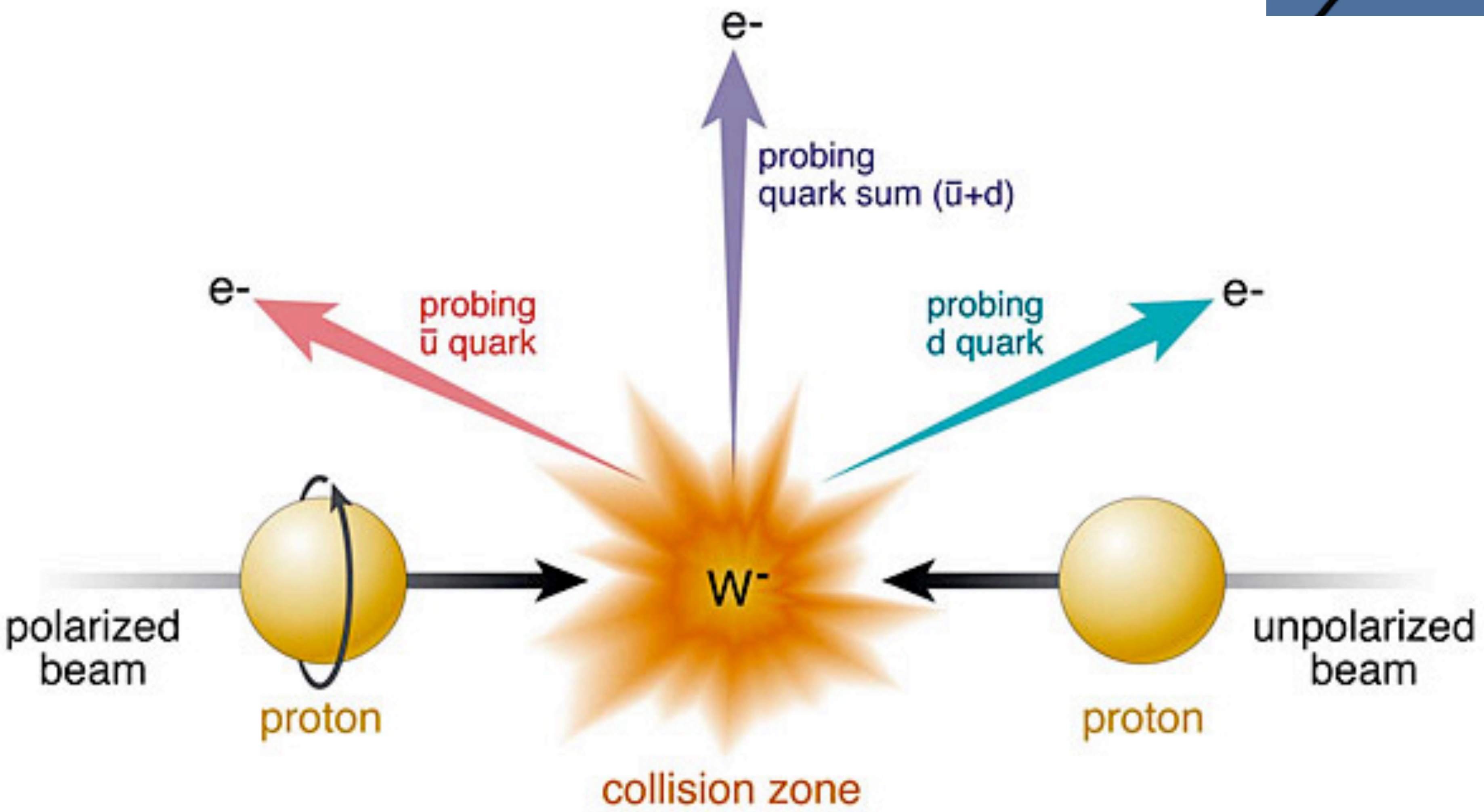


Explore properties of proton spin using W boson

"Exploring the mystery of **proton spin** has been one of the key scientific research goals at RHIC," said **Steven Vigdor, Brookhaven's Associate Laboratory Director** for Nuclear and Particle Physics. .... **The W boson measurements [will help us] ... in quantitative understanding of proton spin structure and dynamics.**"



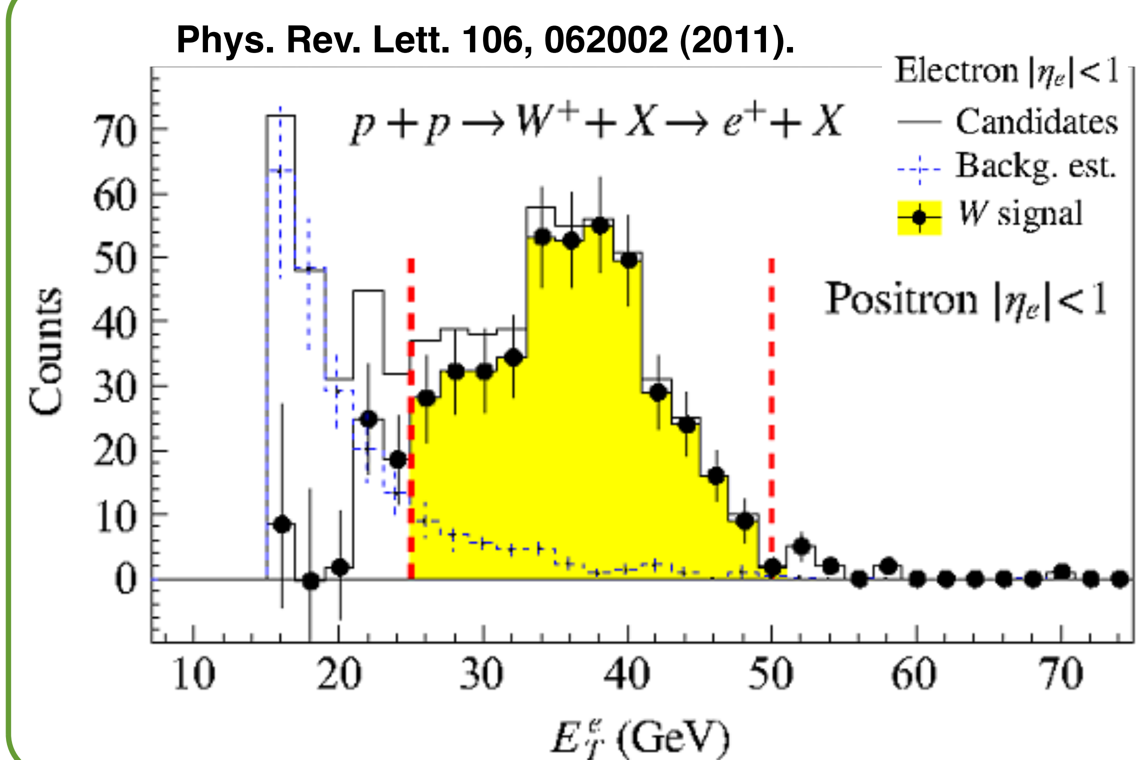
http://www.bnl.gov/bnlweb/pubaf/pr/PR_display.asp?prID=1232

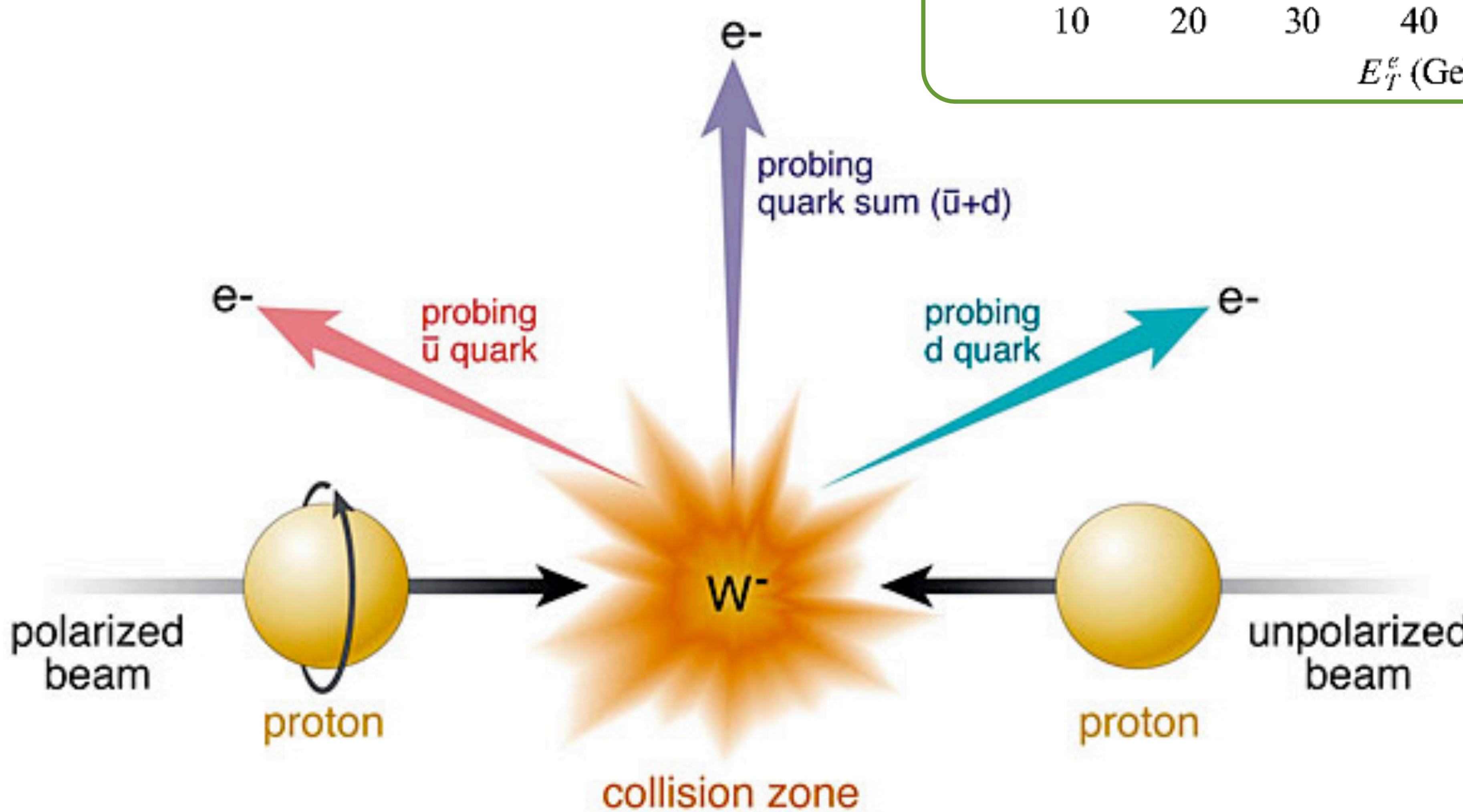


Explore properties of proton spin using W boson

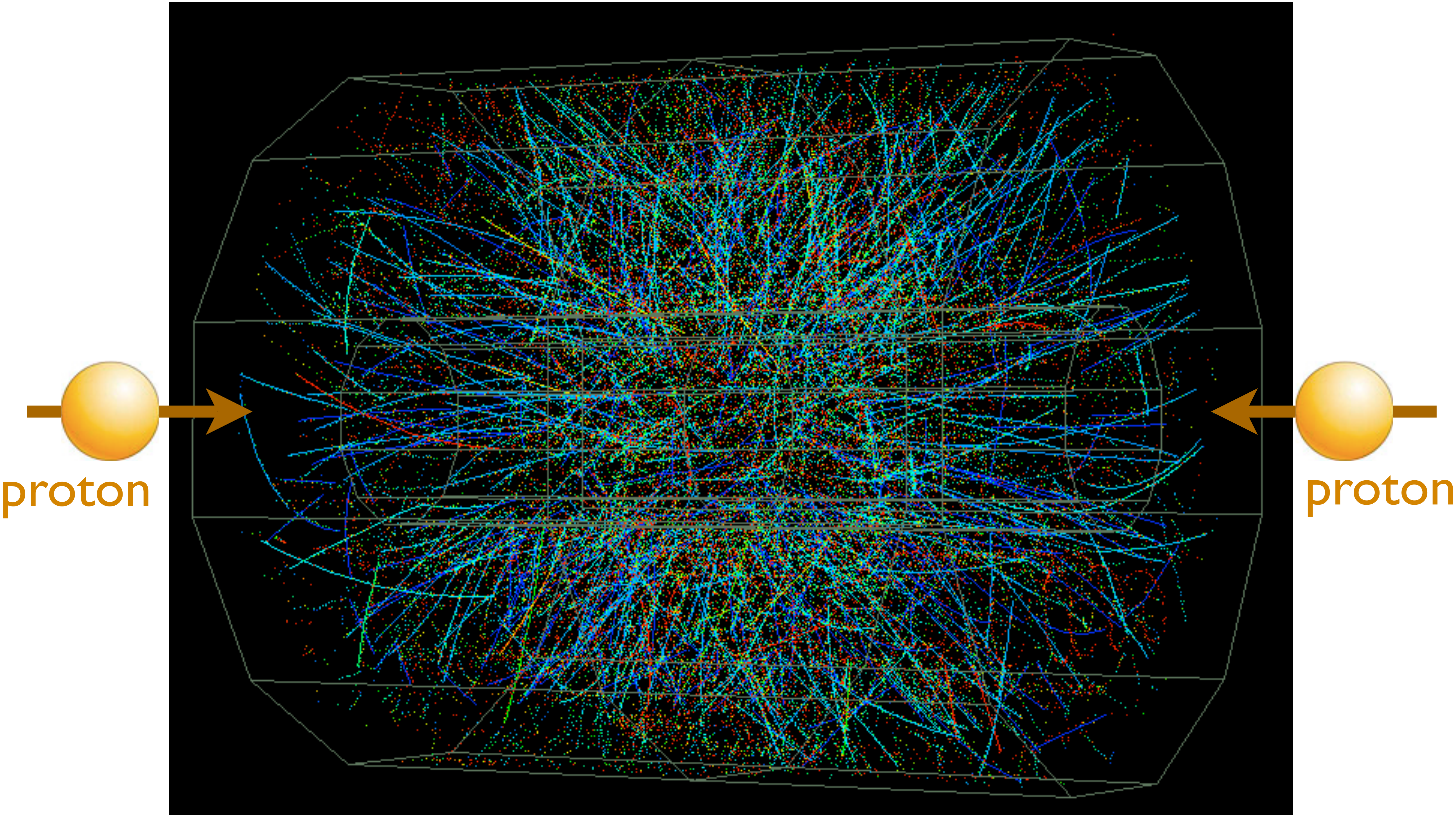
"Exploring the mystery of **proton spin** has been one of the key scientific research goals at RHIC," said **Steven Vigdor, Brookhaven's Associate Laboratory Director** for Nuclear and Particle Physics. .... **The W boson measurements [will help us] ... in quantitative understanding of proton spin structure and dynamics.**"

http://www.bnl.gov/bnlweb/pubaf/pr/PR_display.asp?prID=1232





Registered collision of 2 protons with **lot of energy**



Reconstruction of particles emerging from collision of two protons is a computational challenge

Computational challenges at STAR for W physics

Data Acquisition

- STAR records 'events' at 1kHz
- data rate ~1 GiB/sec

event file: 5 GB with 15,000 events

Data Reconstruction

- reconstruction of 1 event : 10 seconds
- time to process 5GB event file: 40 hours
- **10,000 CPUs needed for a true real time event processing**

Computational challenges at STAR for W physics

Data Acquisition

- STAR records 'events' at 1kHz
- data rate ~1 GiB/sec

event file: 5 GB with 15,000 events

Data Reconstruction

- reconstruction of 1 event : 10 seconds
- time to process 5GB event file: 40 hours
- **10,000 CPUs needed for a true real time event processing**

Analysis requires Calibration of Detector response

- **Quality : crude** , available within an hour

Cloud computing → application

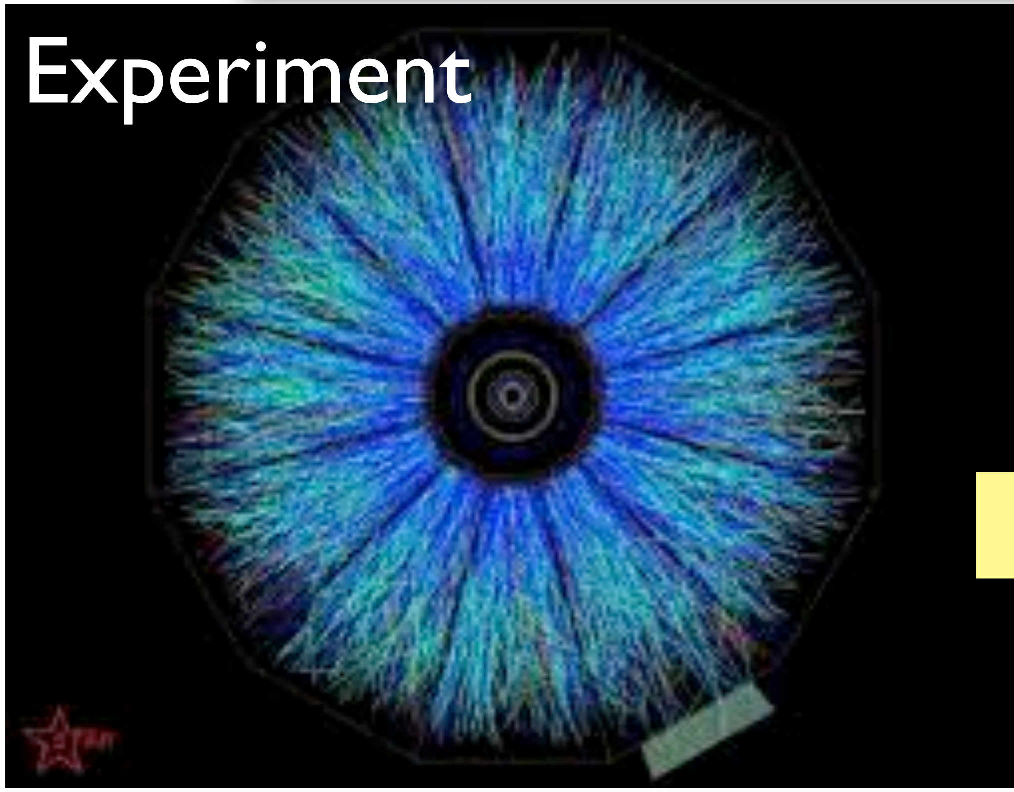
'fastOffline' reconstruction of 15% of events, used to monitor performance of detector

- **Quality : preliminary**, available within a month

  start first data pass
- **Quality : final** , available within 6 months

  full data pass over all qualified events, used for publication of results

Traditional in-house data analysis model

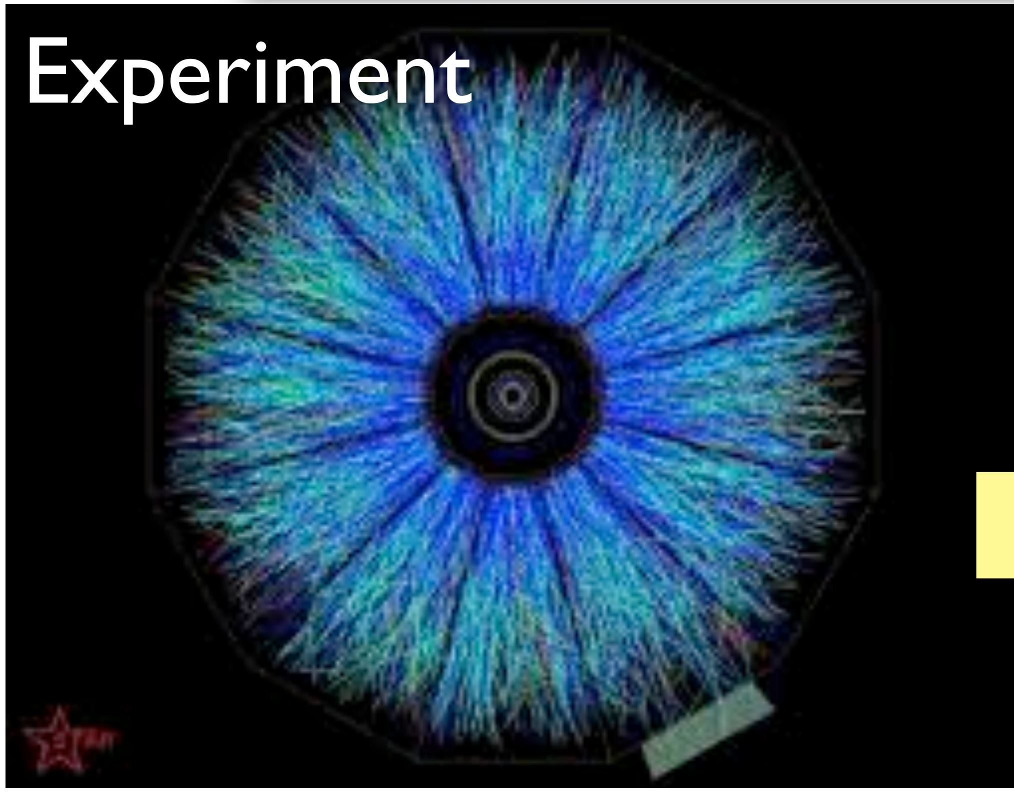


for 3 months

raw data 1 kHz

HPSS

Traditional in-house data analysis model



HPSS

for 3 months

raw data 1 kHz

raw data 300 Hz

for 1 year

results

In-house computing

farm of 2000 dual core machines running highly customize analysis package

Virtualization enables outsourcing of computation

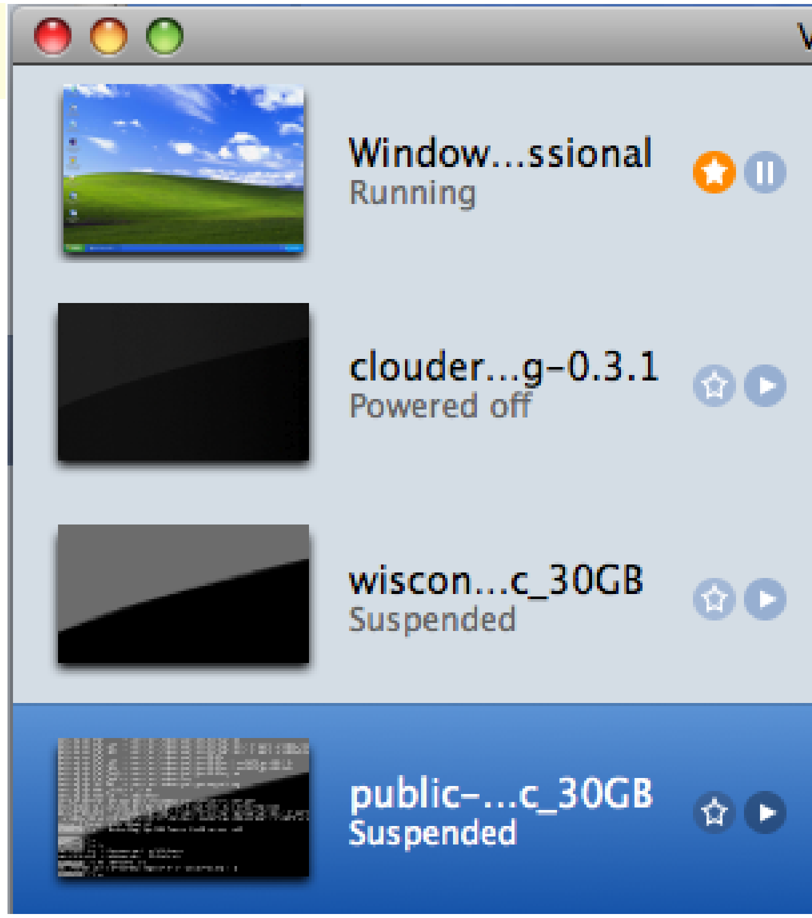
STAR Virtual Machine (VM) is born ...

at first on a laptop ....

Virtualization enables outsourcing of computation

STAR Virtual Machine (VM) is born ...

at first on a laptop ....



1) recently STAR VM is **prepared at a PC** at NERSC

2) pack it 'from inside' and **ship to** Amazon EC2, Magellan@NERSC, Magellan@ANL, etc..

Virtualization enables outsourcing of computation

STAR Virtual Machine (VM) is born ...

at first on a laptop ....

1) recently STAR VM is **prepared at a PC** at NERSC

2) pack it 'from inside' and **ship to** Amazon EC2, Magellan@NERSC, Magellan@ANL, etc..

- Validate once, re-use multiple times.
- The same results obtained ANYWHERE
  → virtualization allows normalization of resources
- Reproducibility of old code results rests in archived old VM, no need to retain hardware

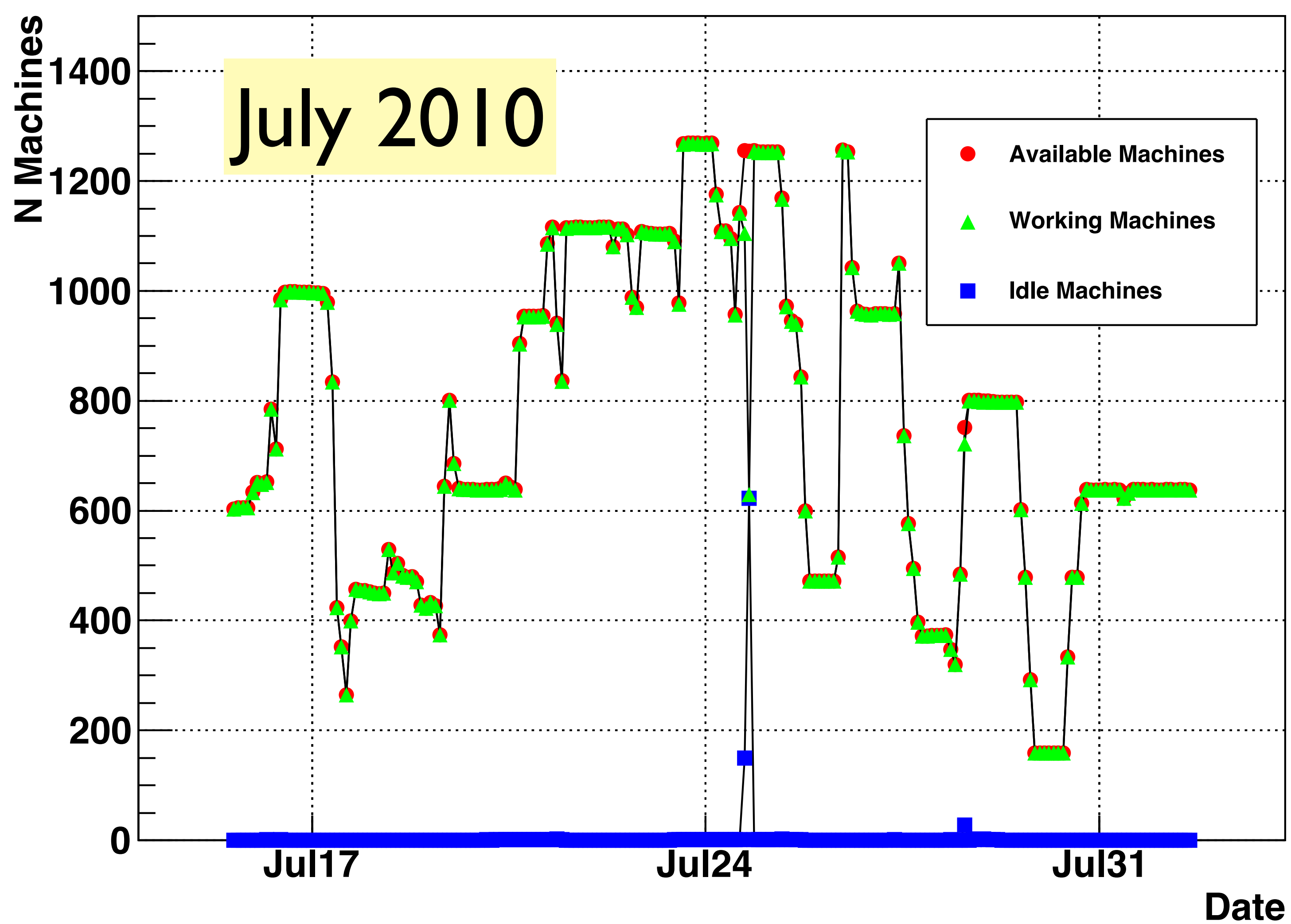
STAR encounters with VMs

date	Facility	tools	type of task	# of VMs	# jobs/ VM	total CPU days	calendar days	total input (TB)	total output (TB)	remarks
2009, March	Amazon EC2	Nimbus Globus PBS batch	simu	100	1	500	5	0	0.3	works like normal globus GK grid site
2009, November	Amazon EC2	EC2	simu	10	1 or 2	1	1	0	0.01	use commercial interface
2010, February	GLOW Madison Uni Wisconsin	CondorVM	simu	430	1	130	0.6	0	0.1	call home model
2010, July	Clemson Uni, SC	Kestrel, QEMU-KVM	simu	1000	1	**17,000**	20	0	7	VM lifetime 24 h, no ssh to VM
2011, February	Magellan NERSC	Eucalyptus	**data reco**	20	6 or 7	600+	20+	2	1	**almost real-time processing**

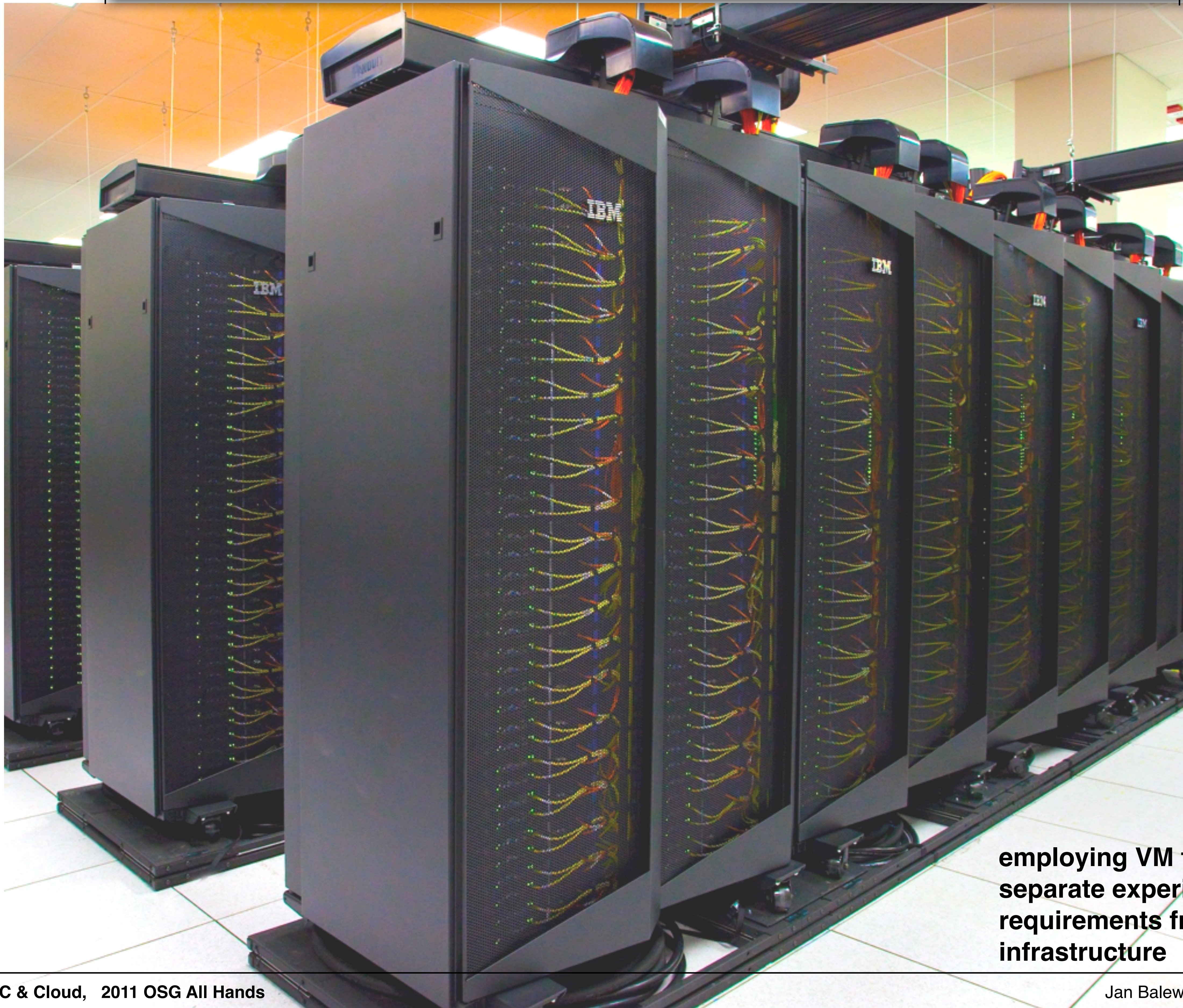
Largest STAR simulations (ever) at Clemson

- STAR MC simulations with partonic p_T > 2 GeV, PYTHIA event generator
- **Virtual Machine** prepared with STAR software stack and **deployed to over 1000 machines**
- Using cloud computing at **Clemson University in South Carolina** (Ranked #85 best supercomputer)

- Over 12 billion events generated
- Took over **400,000 CPU hours** and generated **7 TB of data** transferred to BNL
- Largest physics simulation on cloud, largest STAR simulation in CPU hours
- **Benefit: shorten by a year PhD study of MIT student**

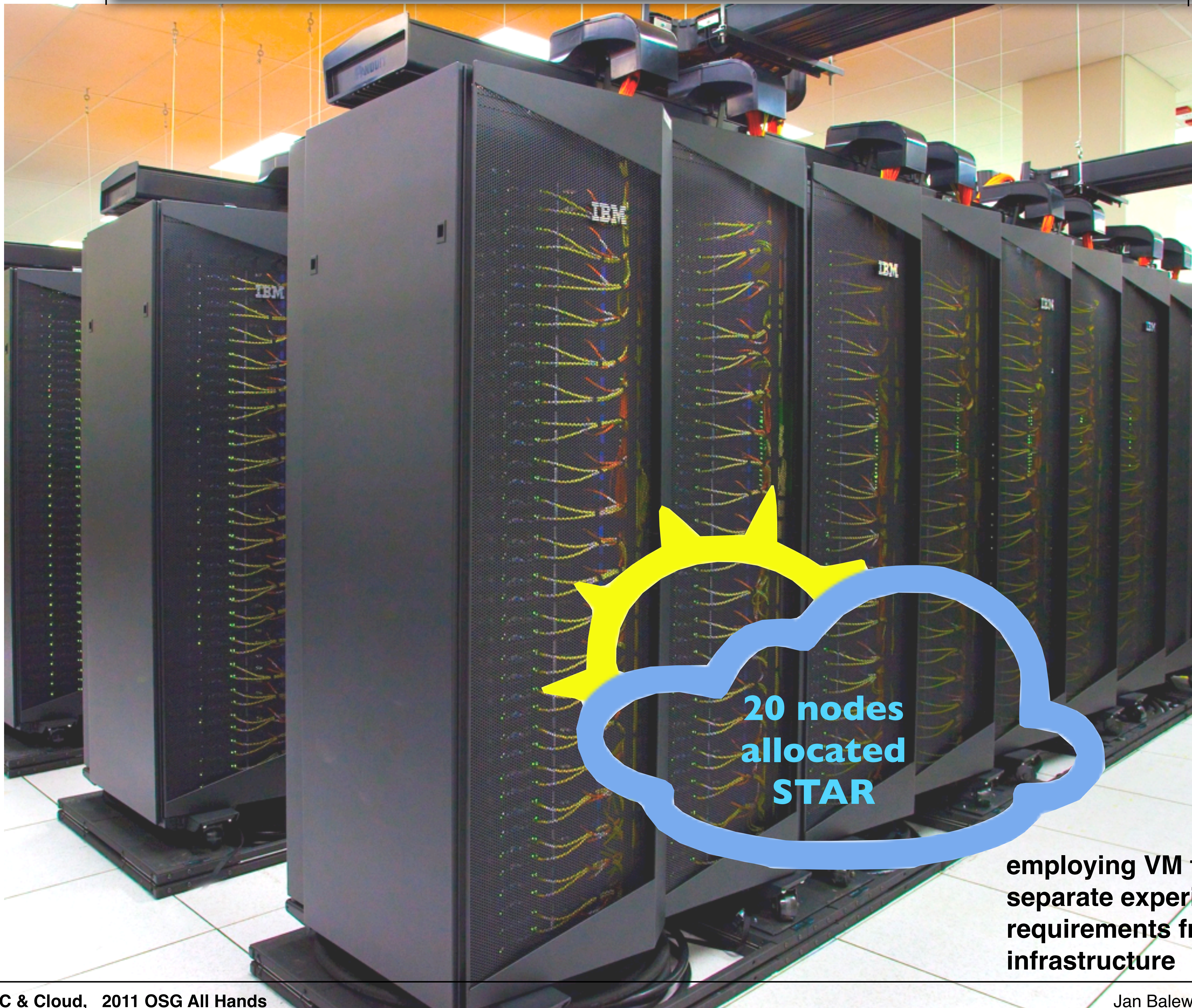


Today: Magellan @ NERSC



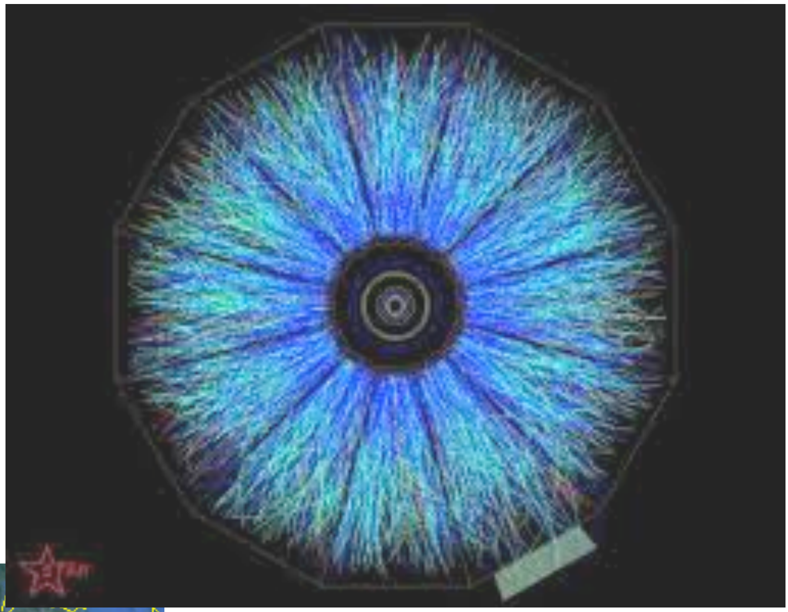
employing VM technology to separate experiment specific requirements from facility infrastructure

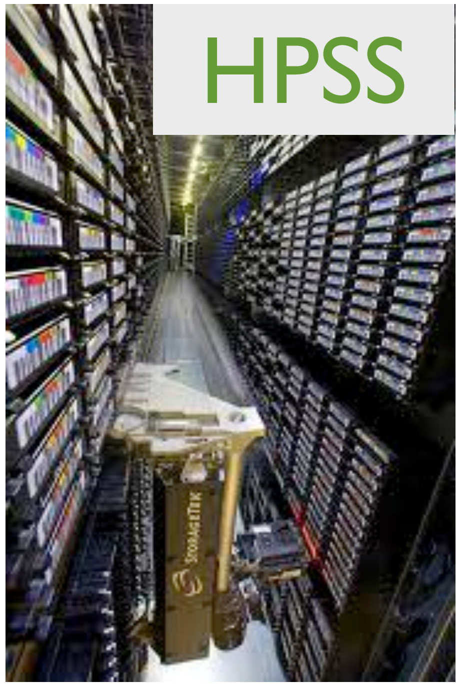
Today: Magellan @ NERSC



employing VM technology to separate experiment specific requirements from facility infrastructure

STAR
experiment
@BNL



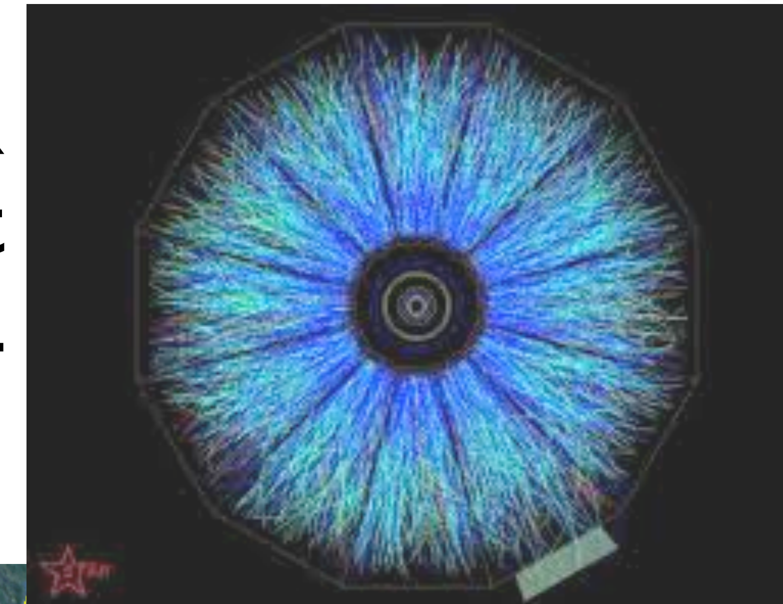


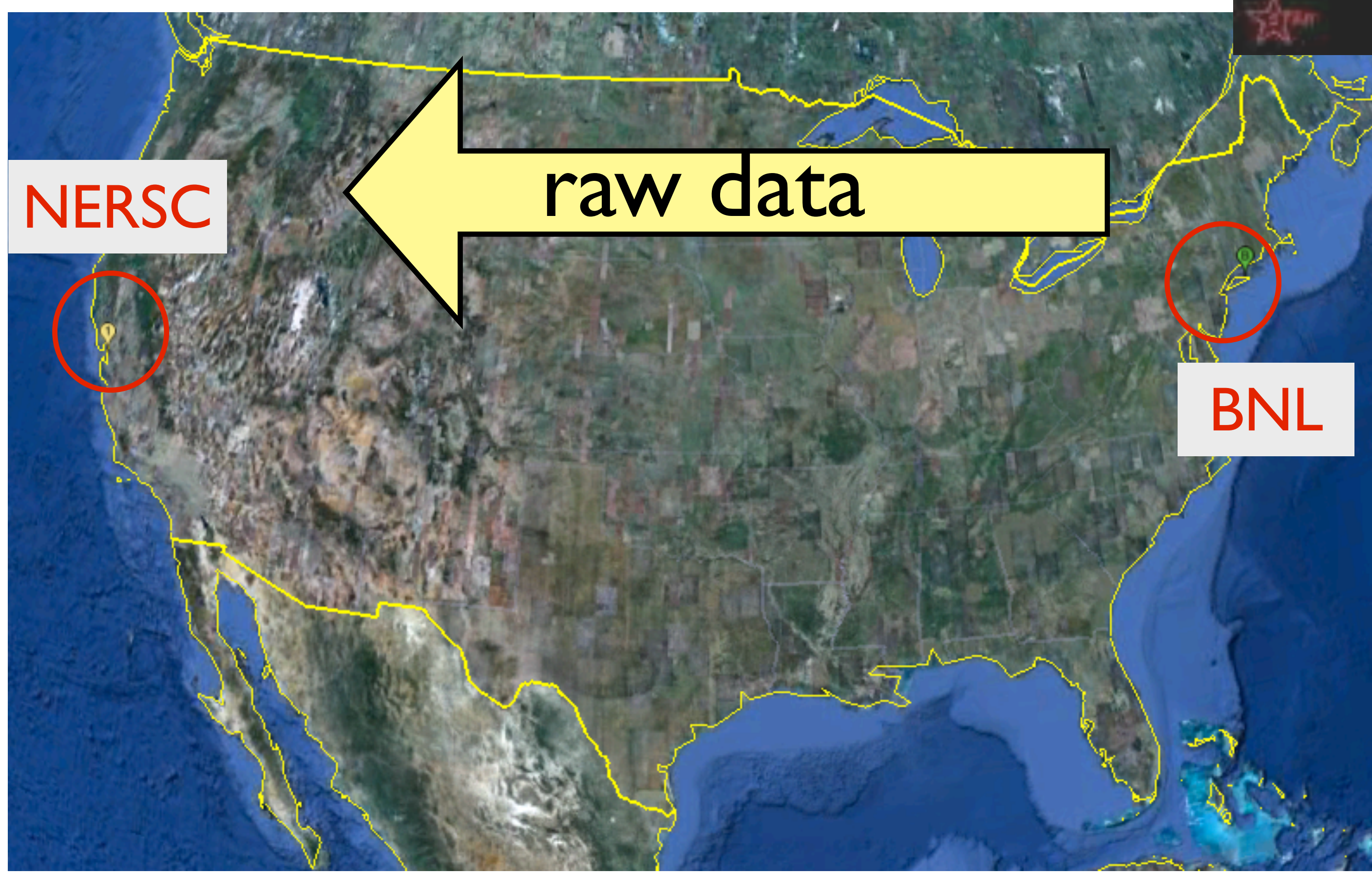
Real-time distributed processing of 2011 Data

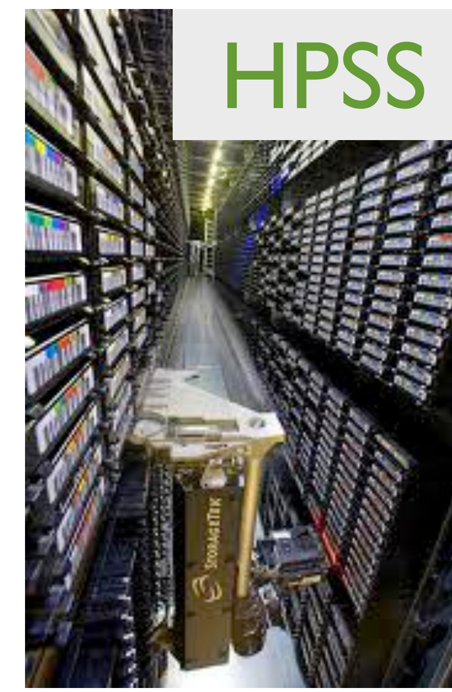
clone
STAR VM
x 20

Magellan @
NERSC

STAR
experiment
@BNL





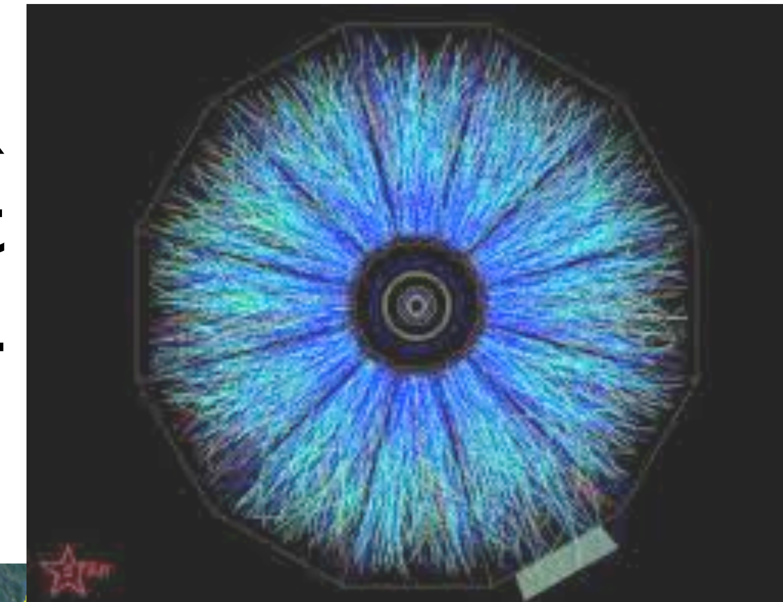


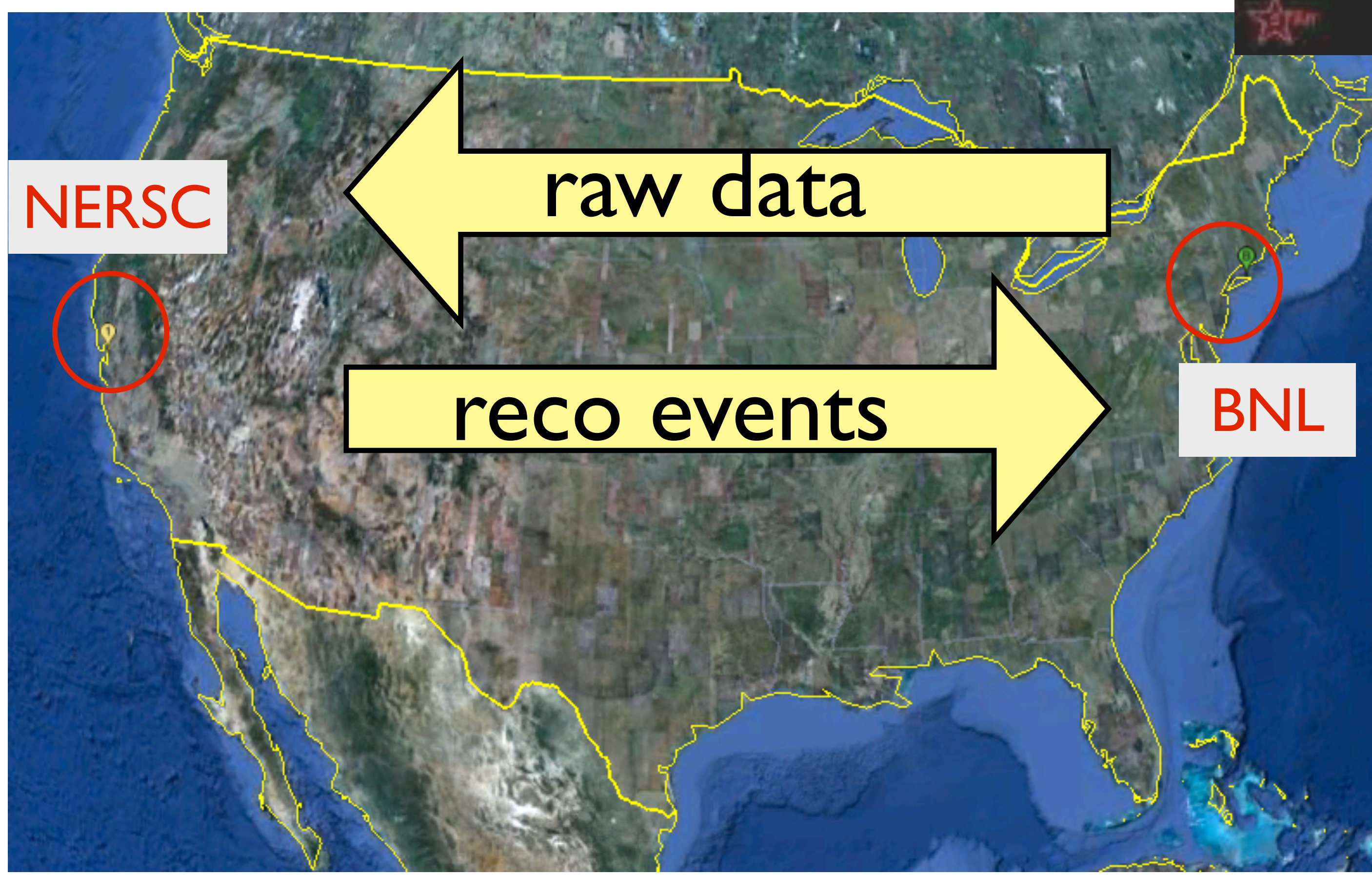
Real-time distributed processing of 2011 Data

clone
STAR VM
x 20

Magellan @
NERSC

STAR
experiment
@BNL



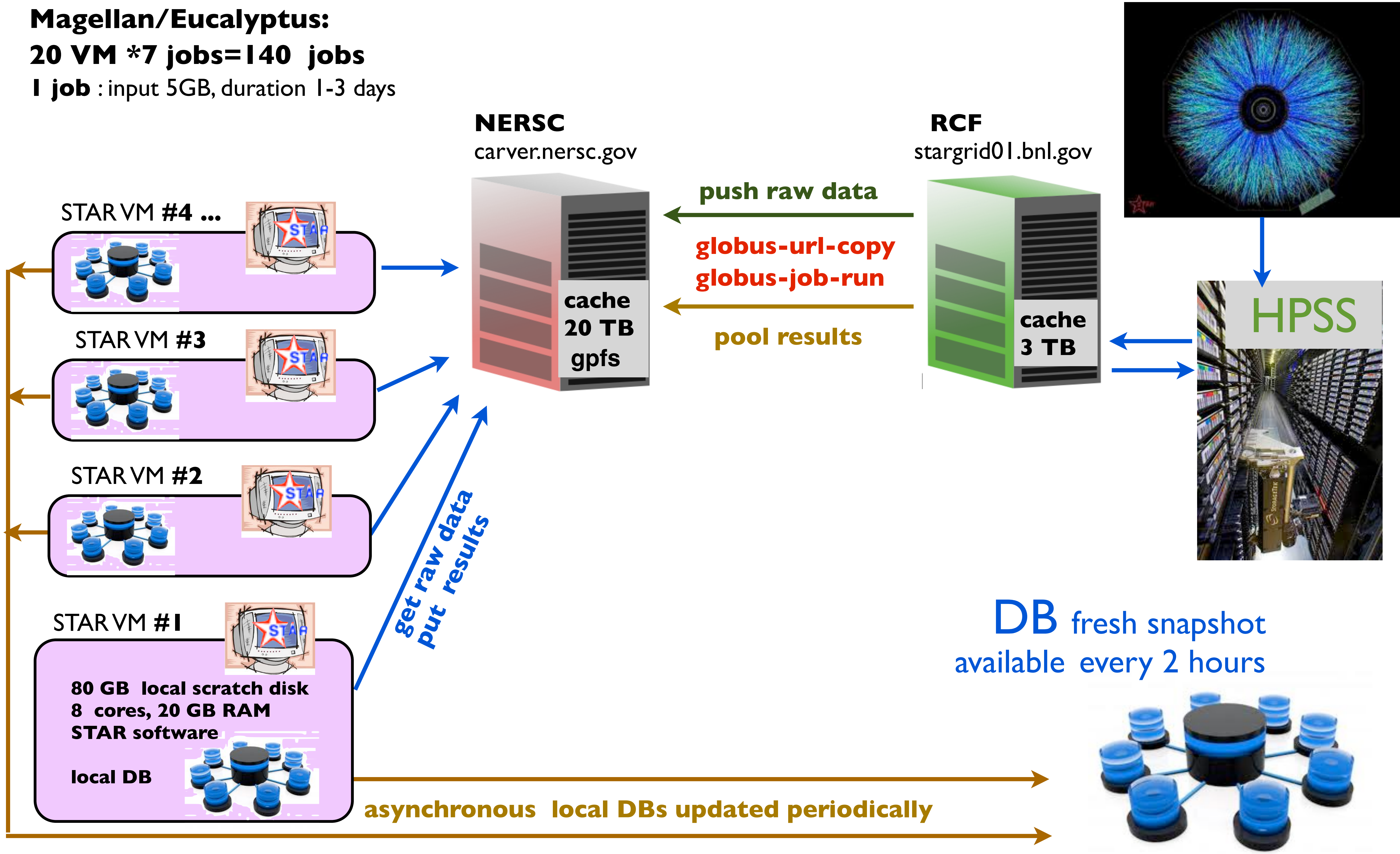


DB
w/ time dependent
detector calibration

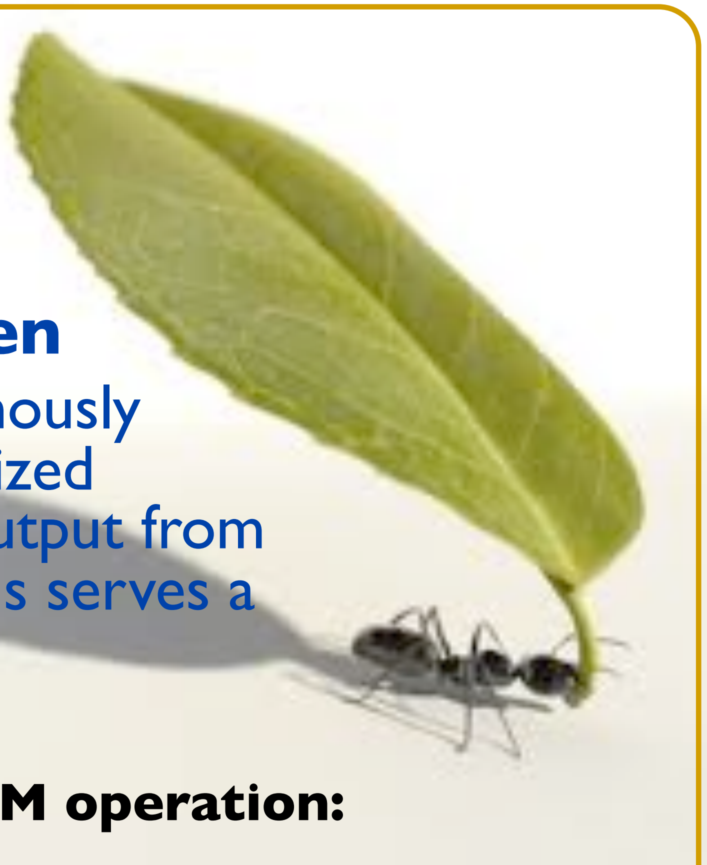
Topology of connectivity RCF ↔ VMs

RCF @ BNL

Magellan/Eucalyptus:
20 VM *7 jobs=140 jobs
1 job : input 5GB, duration 1-3 days



Model of coordination of VMs



Model citizen

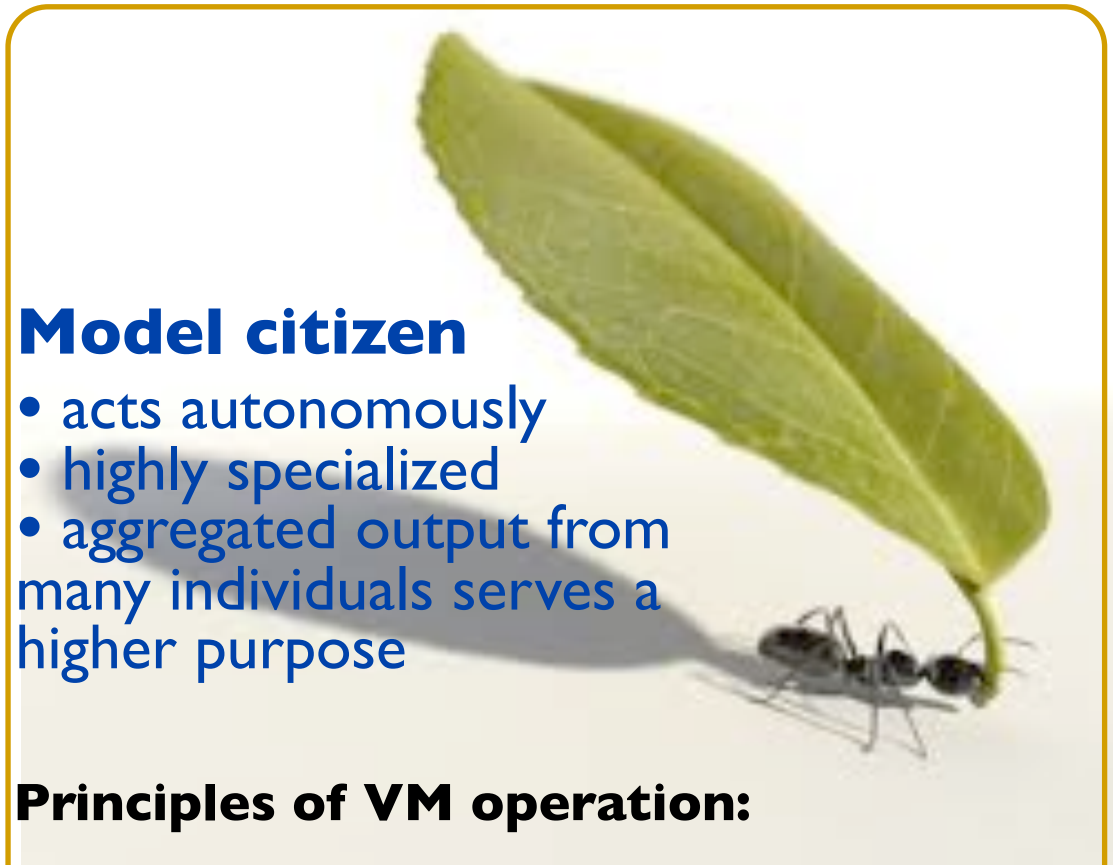
- acts autonomously
- highly specialized
- aggregated output from many individuals serves a higher purpose

Principles of VM operation:

1. Acts w/o supervision
2. Protects own integrity
3. Initiates connection to host
   - acquire input
   - perform task
   - retruns results to host
   - rest for '5 minutes'

pagoda nest-ants nest

Model of coordination of VMs



Model citizen

- acts autonomously
- highly specialized
- aggregated output from many individuals serves a higher purpose

Principles of VM operation:

1. Acts w/o supervision
2. Protects own integrity
3. Initiates connection to host
   - acquire input
   - perform task
   - retruns results to host
   - rest for '5 minutes'

New job will start only if VM **state=open**

VM
8 cores, 20 GB RAM, 80 GB disk
local mySql DB

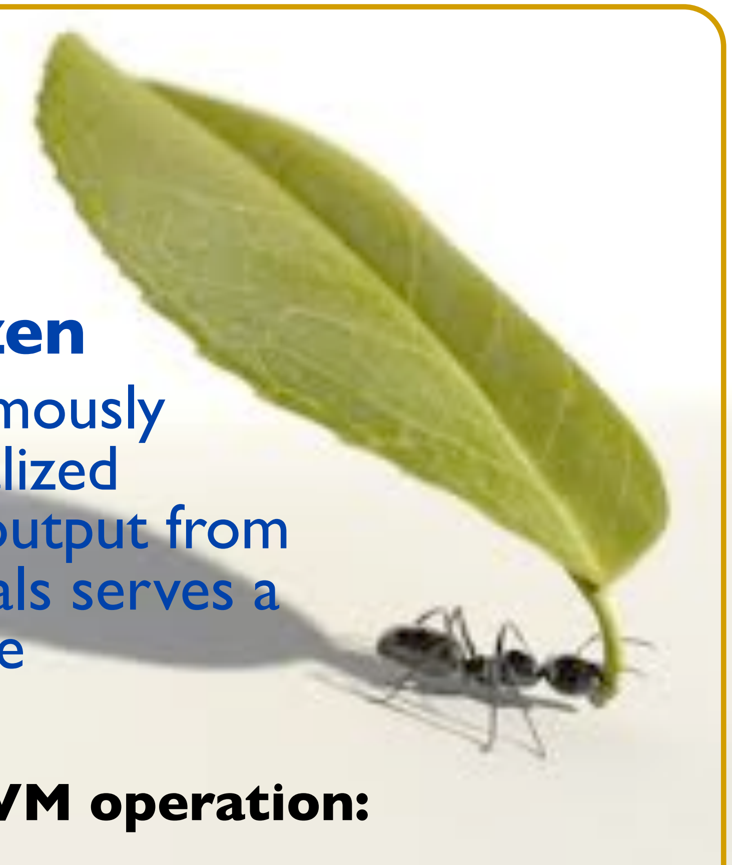
self-check (state machine)

case:

load >10	→ **hotCPU**
disk<35 GB	→ **diskFull**
# jobs>6	→ **busy**
DB too old	→ **oldDB**
jobs enabled	→ **open**
default:	→ **closed**

Model of coordination of VMs

New job will start only if VM **state=open**



Model citizen
- acts autonomously
- highly specialized
- aggregated output from many individuals serves a higher purpose

Principles of VM operation:

1. Acts w/o supervision
2. Protects own integrity
3. Initiates connection to host
   - acquire input
   - perform task
   - retruns results to host
   - rest for '5 minutes'

VM
8 cores, 20 GB RAM, 80 GB disk
local mySql DB

self-check (state machine)

case:

load >10	→	**hotCPU**
disk<35 GB	→	**diskFull**
# jobs>6	→	**busy**
DB too old	→	**oldDB**
jobs enabled	→	**open**
default:	→	**closed**

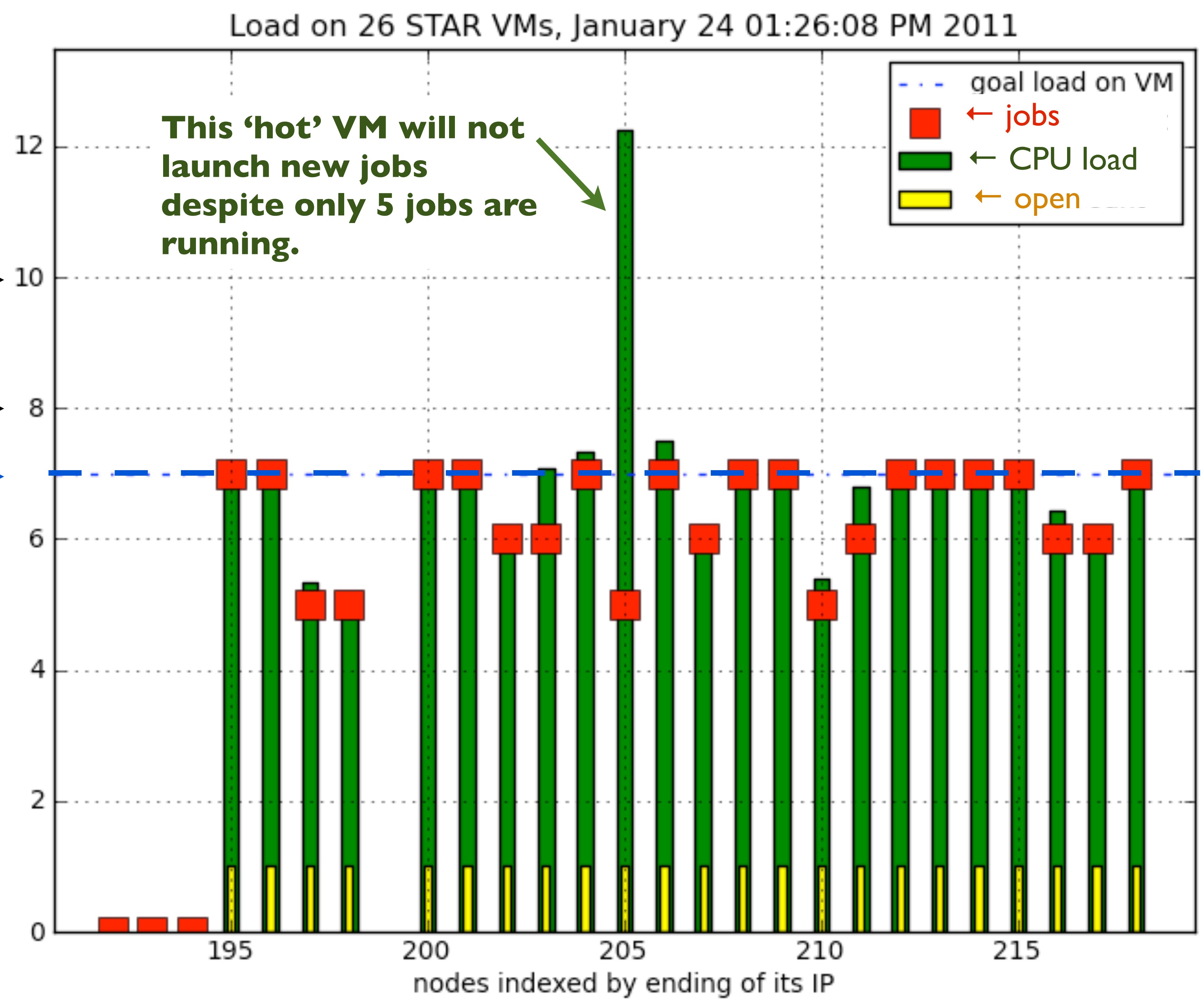
- No micro-management of instance, local rules result with coherent aggregated output
- No active reporting by VMs
- No inter-machines synchronization
- VMs compete for data
  use of 'atomic' rename avoids collisions
- Instances are disposable and failures don't disrupt workload of other N-1

VM cluster load - example

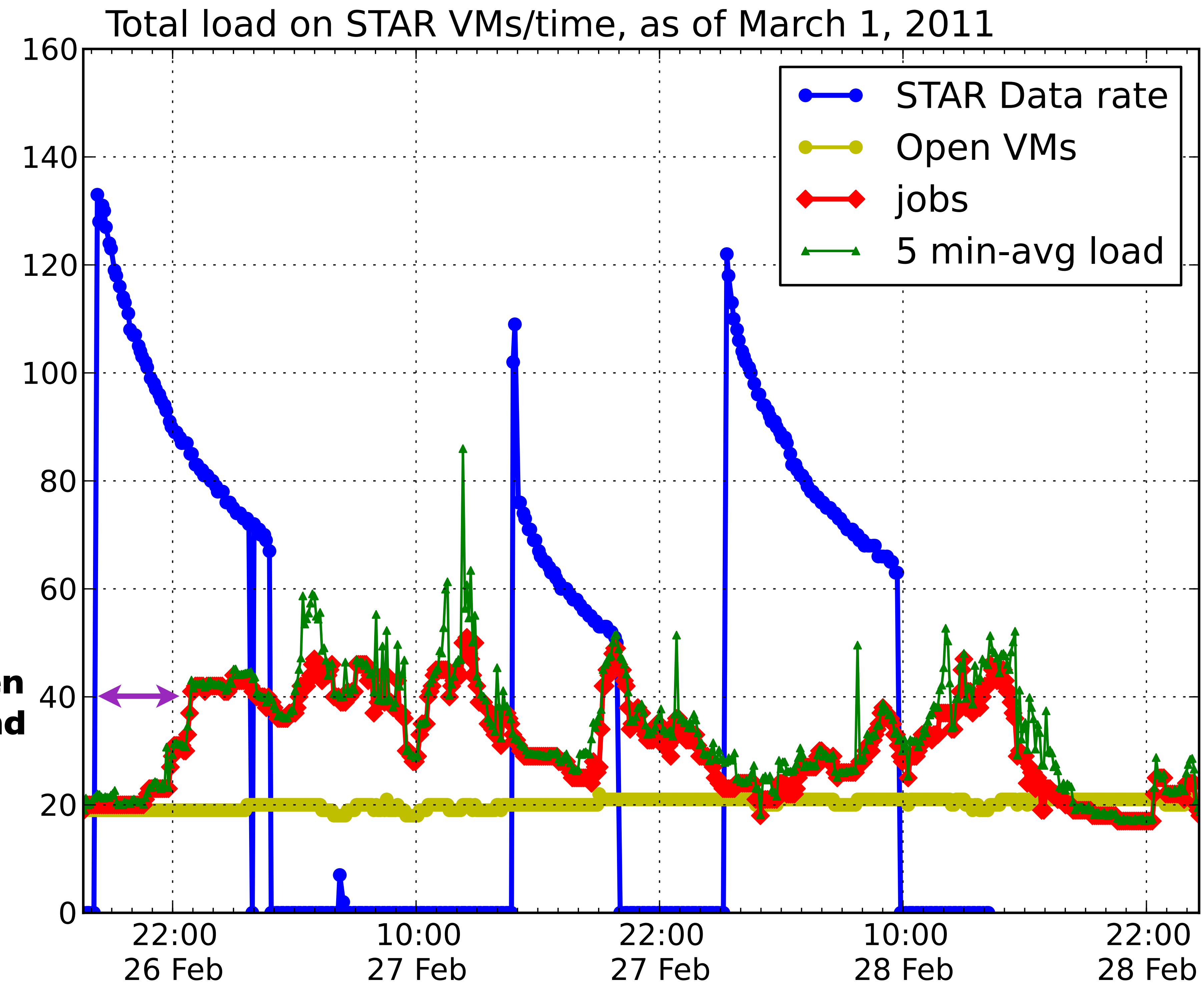
'hot VM' threshold →

# of cores per VM →

goal # of jobs →



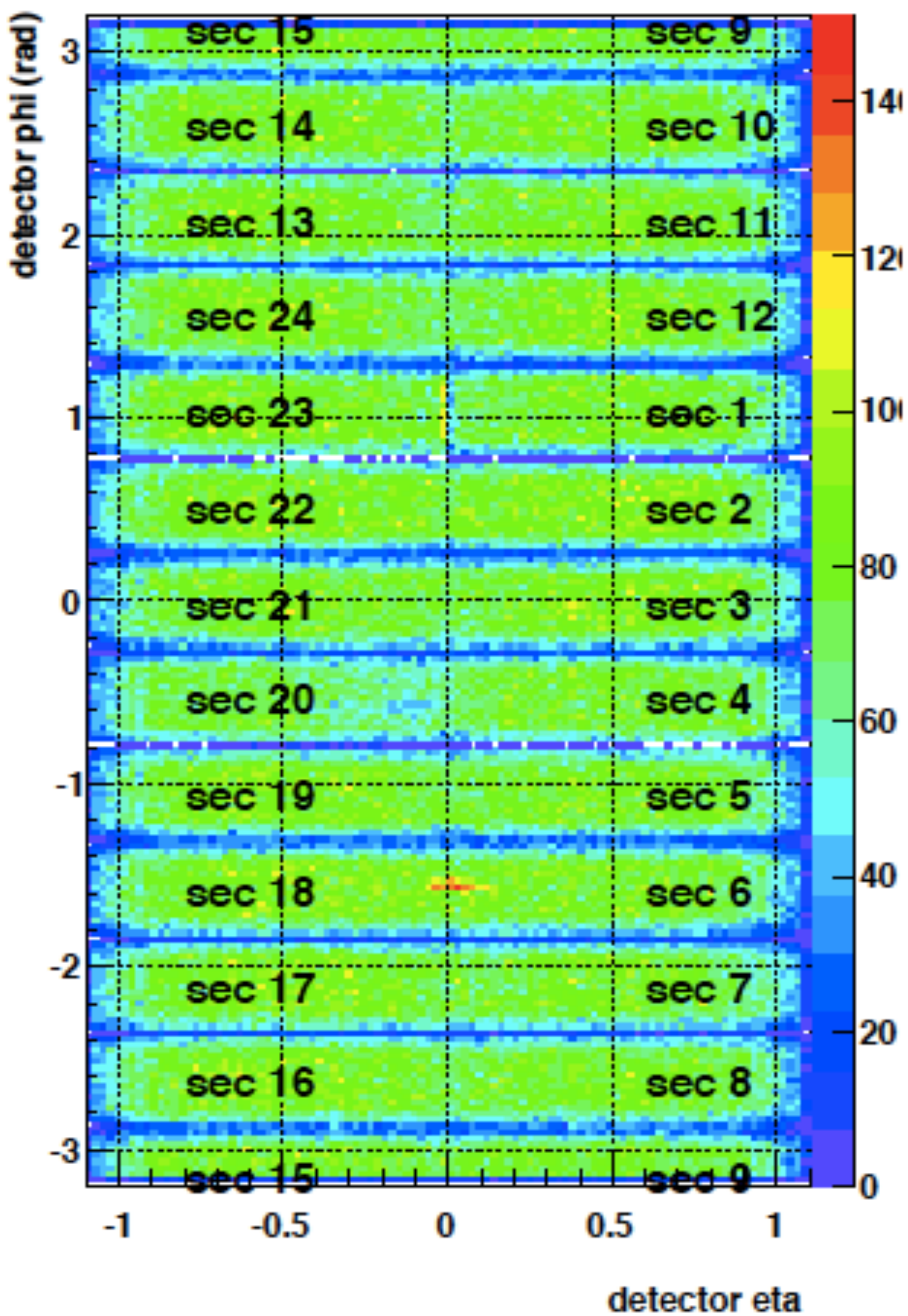
Time lag of 'real-time' processing



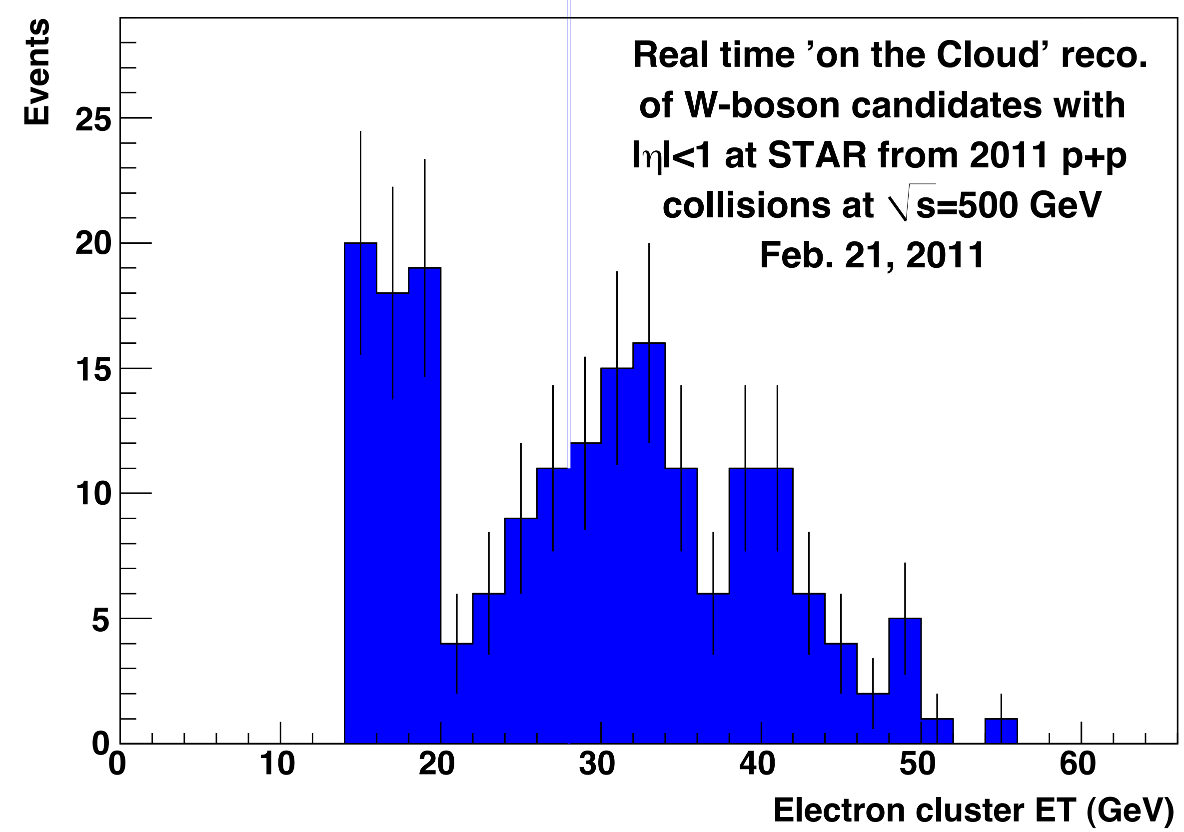
few hours latency between data acquisition (blue) and reconstruction (red)

Deliverables after 10 days of data taking

Uniformity of reconstructed tracks in 2011 data

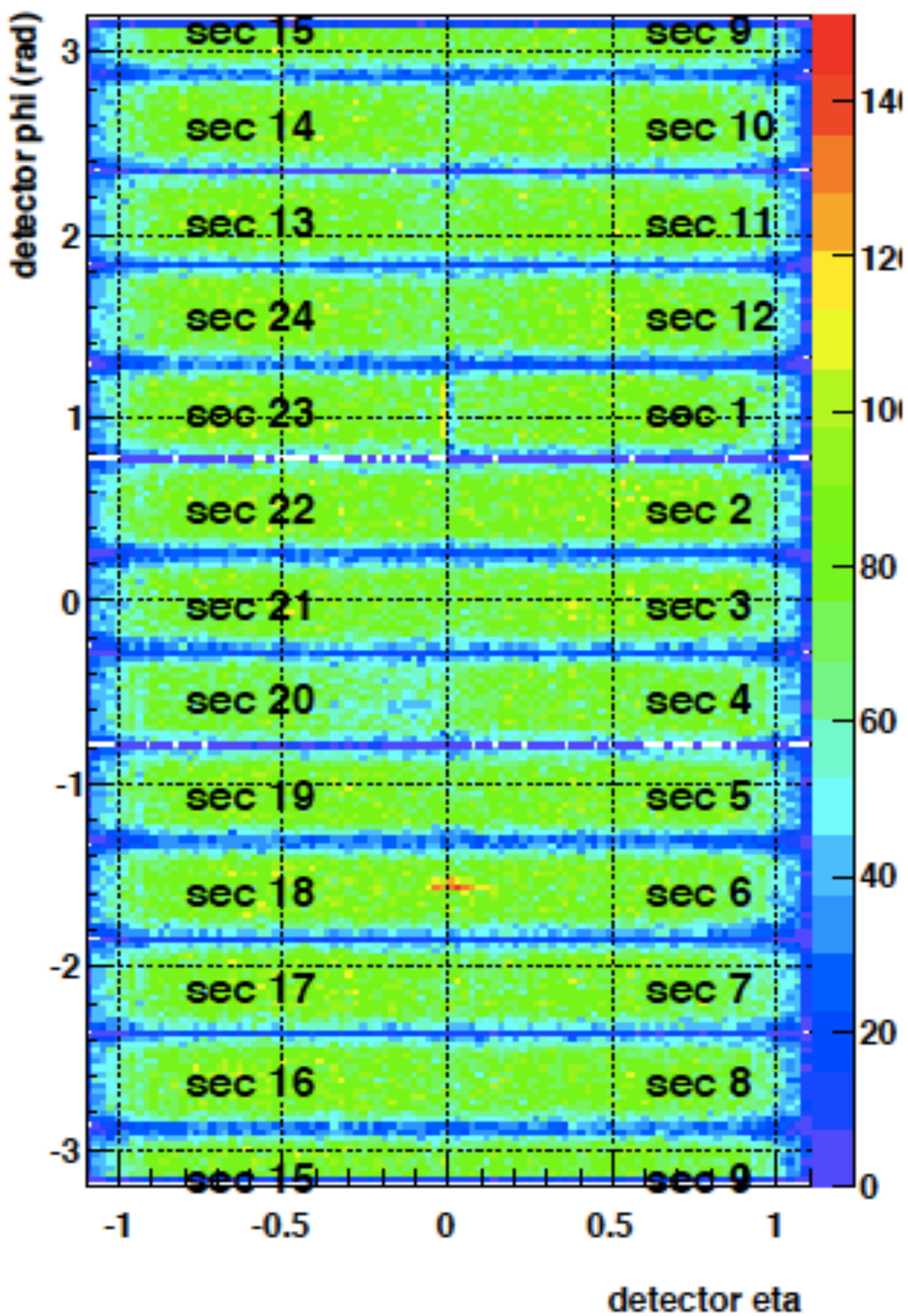


STAR first 100 Ws reconstructed in 2011 using Cloud resources: Magellan @ NERSC

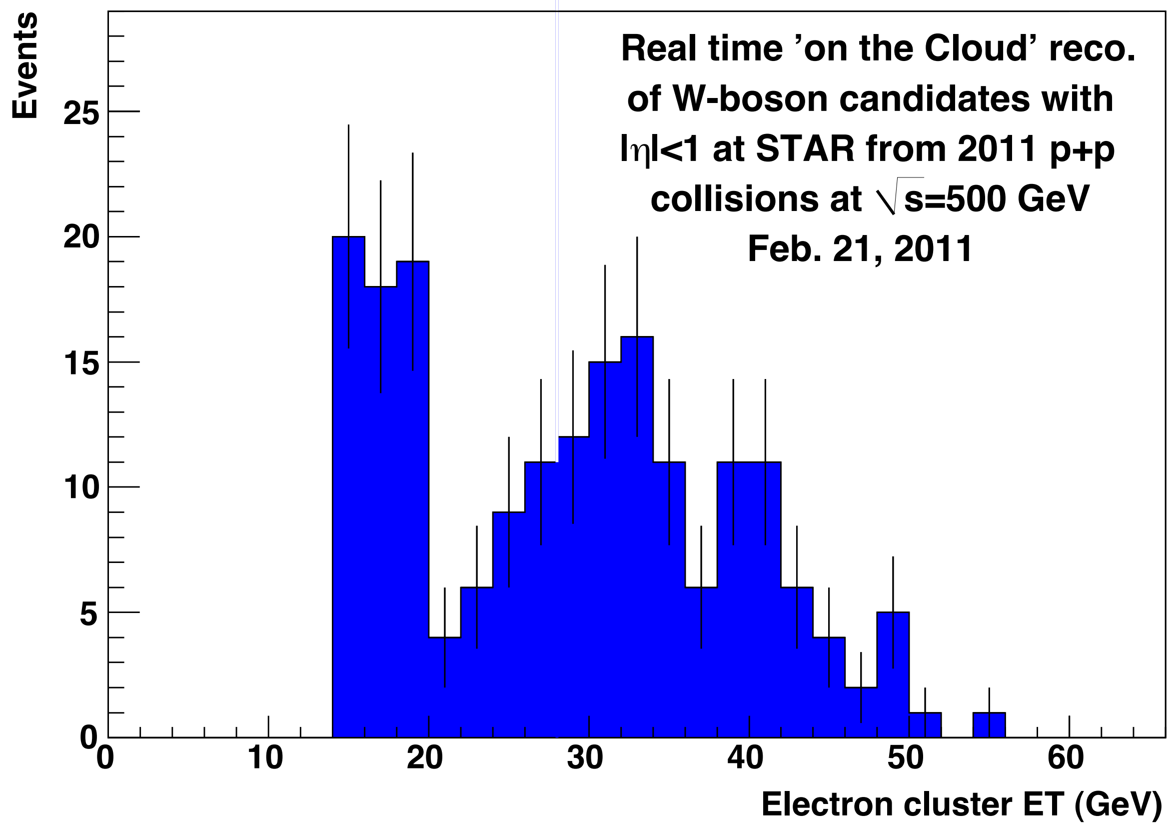


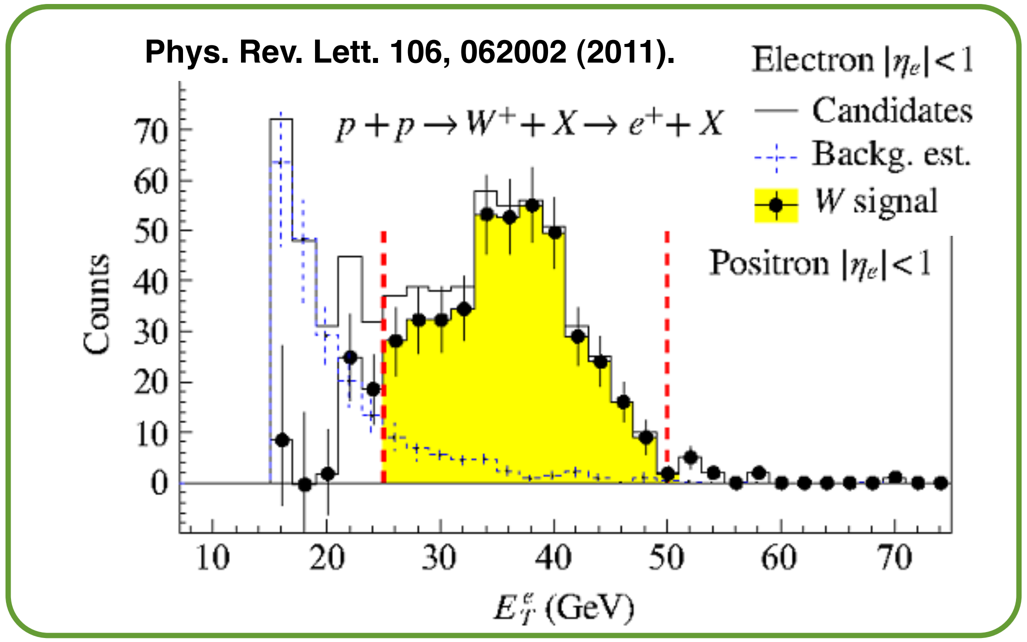
Deliverables after 10 days of data taking

Uniformity of reconstructed tracks in 2011 data



STAR first 100 Ws reconstructed in 2011 using Cloud resources: Magellan @ NERSC



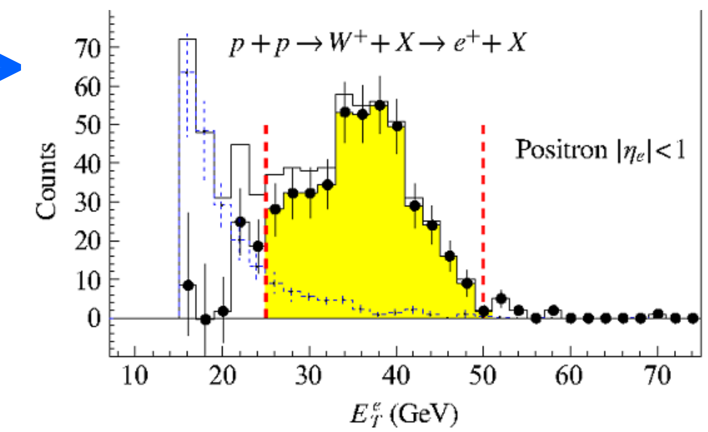


How much of ACCESS is SUCCESS ?

Achieved timeline of W measurement in 2009

	2009 →										2010→				
	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
data taking															
calibration															
reco pass 1													presentation on		
analysis													☆ a conference		

scientific deliverables in 10 months

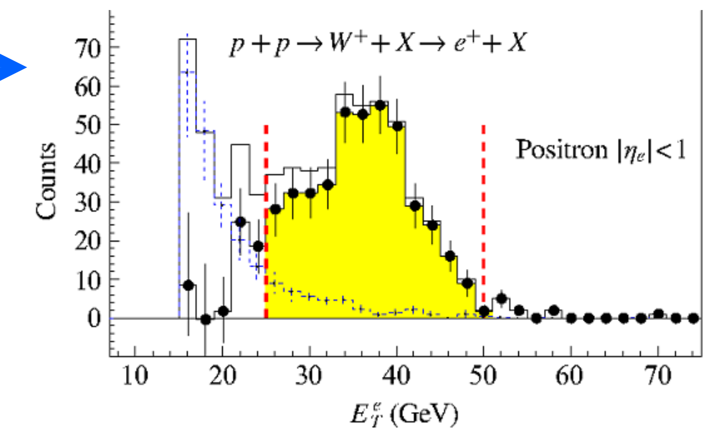


How much of ACCESS is SUCCESS ?

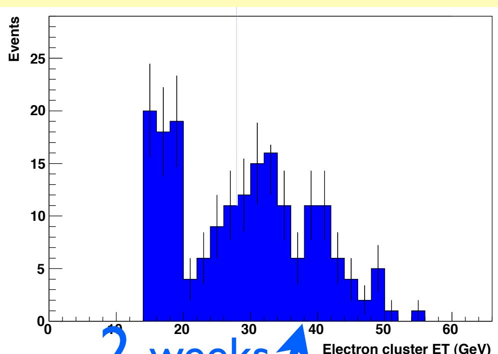
Achieved timeline of W measurement in 2009

	2009 →										2010→				
	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
data taking															
calibration															
reconstruction													presentation on		
analysis													☆ a conference		

scientific deliverables in 10 months



Intended timeline of W measurement in 2011



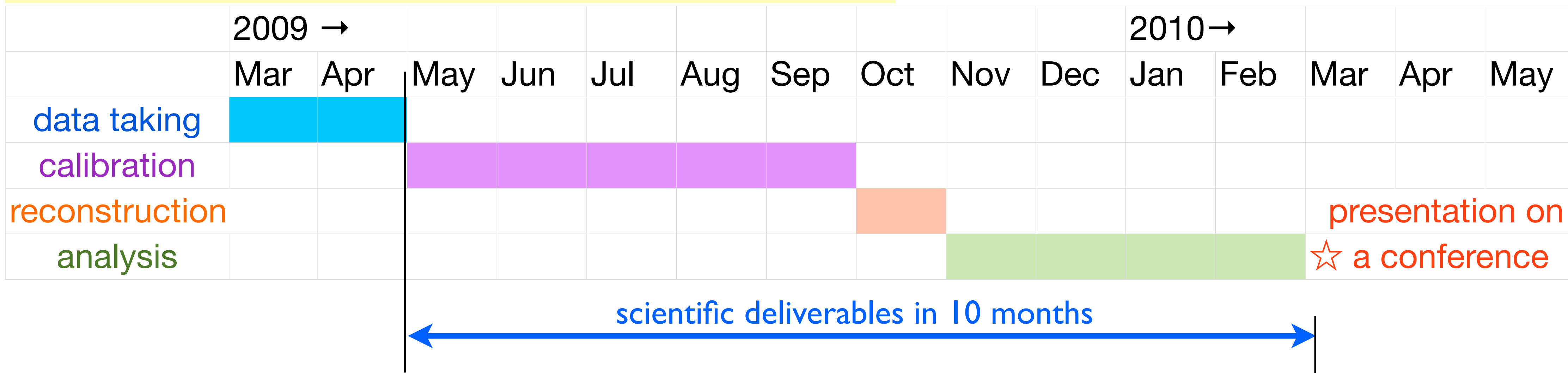
2 weeks

deliverables in 3 months ?

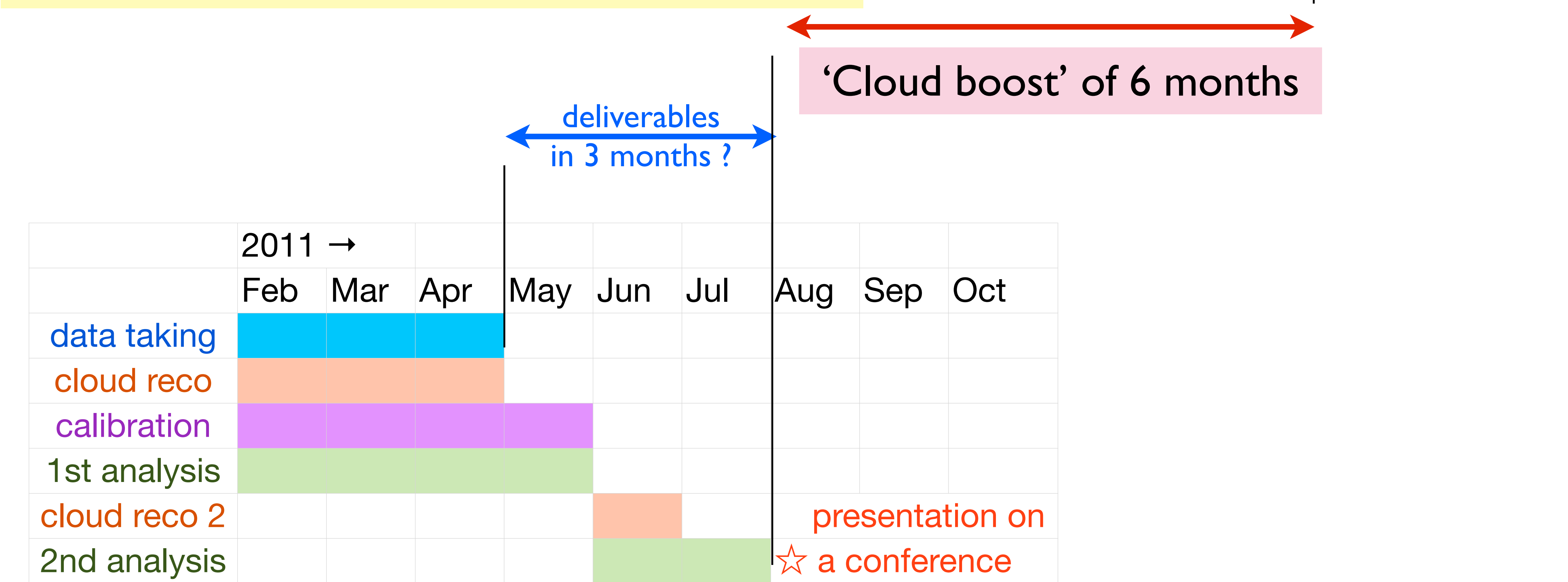
	2011 →								
	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
data taking									
cloud reco									
calibration									
1st analysis									
cloud reco 2							presentation on		
2nd analysis							☆ a conference		

How much of ACCESS is SUCCESS ?

Achieved timeline of W measurement in 2009



Intended timeline of W measurement in 2011



Summary

- Virtualization has allowed STAR to run complex workflow and address intense processing demands

- Today STAR is doing real data reconstruction in near real time, providing a valuable QA and preview of the results

  (preliminaries for the W to be used for making the Physics case far ahead of final publishable results)

- Processing on a distributed facility are real and beyond proof of principles

  (STAR is doing this TODAY at a 7% level - scalability ramp up is next)

- Availability of such capabilities in OSG would allow full exploitation of resources available on a distributed National Facility

- **Thanks to virtualization capabilities of Cloud** and the resources provided by the Magellan project and the Magellan support team at NERSC **STAR is in a world-wide unique position to process acquired data in real time.** Experimentalist can see what they measure as they measure.

- Faster data analysis will shorten publication cycle

- Unified VMs allow easy integration over multiple geographical location

BACKUP

The real 'thing' can be seen here:

http://portal.nersc.gov/project/star/balewski/w2011/C/

Multi-site real time STAR data reco , March 2011