Comments on #3:
“Motivation for Regional Analysis Centers and Use Cases”

Chip Brock

3/13/2

what's the problem?

well, no *problem*, really

- an offsite architecture has lots of components and correlations
  - and it's complicated - it has to mesh with the existing SAM, GRID, and database group efforts
  - how do we design it?
  - how do we explain it?
- one way to approach it's design is from a systems perspective
- another way is from specifically how it will be used in practice
  - that's what I'm trying to do
  - I suspect that through illustrations - what I've been calling "stories" - we can head off questions and help to focus ourselves
    - seems a luxury to have this planning opportunity now!

what's a story?

pick a small set of measurements/tasks

- imagine what each step of such a measurement would be
- do it in the context of the data formats that we have available
- try to design the offsite architecture to accommodate these sorts of tasks
- do it in the context likely kinds of offsite institutions

I picked three imagined tasks (yesterday)

- each exercises a different part of the data tier

1. measurement of the W cross section
2. establishing a jet cone energy algorithm
3. determining the em scale corrections

data formats

As I understand it, these are the data formats:

- ROOT tuples
- thumbnail

  5kB, clusters, named trigger, muon hits, jets, taus, vertices
- DST

  50kB, resident on disk, standard RECO output, hits, clusters, global tracks, non-tracking raw data, contains TMB
- DBG

  500kB, created on demand, trigger, cal corrected hits, clusters, vertices, physics objects
- RAW

  250kB, tape

here's what you can maybe do:

DATA FORMAT	TASKS
ROOT TUPLE	
	distributions/fitting
	cross sections
	data-MC comparison
	data-background comparison
	pickED events?
TMB (2)	
5Kb	event display/scanning
em clusters,centr,	root tuple production
named trigger	creation of event lists w/cuts
muon hits,±p	root tuple production
jets	some reconstruction
taus	
MET	
fit track	
vertices	
presh clusters	
DST (1)	
50Kb	DST creation
standard RECO output	event display/scanning
hits	recon code development
clusters	track fitting
gl tracks	track matching algorithms?
non-tracking raw data	jet scale - cone level
	vertexing algorithms
	em scale calibration
DBG (1)	
500Kb	RECO production
created on demand	RECO code development
trigger	channel level calibrations
cal corrected hits	track fitting
clusters	track matching development
track clusters	RECO verification
vertices	channel level calor calib
physics objects	vertex algorithms
	MC production

this is all very sketchy and stream of consciousness now…

just to give an idea of what I think it would be worth putting down

institutions:

I can imagine institutions of generically the following sorts:

US A
faculty, new students only
US B1
faculty, post docs, students
US B2
B1 + RAC/tier II
US C
laboratory
EU A
faculty, post docs, students
EU B
EU A + RAC
EU C
laboratory

where tasks requiring the following data formats might be doable:

	US A	US B1	US B2	US C	EU A	EU B	EU C
ROOT	x 1	x	x	x	x	x	x
TMB		x	x	x	x	x	x
DST		x 2	x	x	x	x	x
DBG			x 3	x		x	x
RAW			x	x			x

em scale cal

jet cone algorithm

W cross section

so, the telling of the story

what's involved in a particular measurement

eg, W cross section:

- nominally: acquire a data set of measured "W's"
- determine/apply corrections
- determine/calculate background
- subtract and count corrected W's
- multiply by live luminosity

in practice:

sitting at your ROOT-capable remote desk:

1.prepare dataset

- how? stream DST's? to an RAC? to home? at FNAL?
- extract them into private root tuples, at RAC? at FNAL?

2.make cuts

- do cuts on root tuples? propagate the cut list info back for any db/DST information? how? SAM? file list? MC analyzed, where? a third site?

3.measure/calculate bckgnds

- where? at ROOT level? prepared where? same as 1-2?

luminosity determined

- where, how? event list, run list prepared from 1., sent back as query to database..where? FNAL? RAC? flat file, cache, subset db prepared and sent where? RAC?

goal of this exercise would be:

describe the ideal procedure

- in terms of real projects
- imagining the real flow of requests, data movement, etc.

If we can do this

- then we know how to design the system
- and we know how to explain it.