the Regional Analysis Center concept

~~5~~ 11 months later

This was an attempt to stimulate a general discussion within the DØ task force charged with evaluating the RAC concept.

The task force is basically done: version 14 of its recommendations appears to be the final word.

The details here are pretty close to those recommendations.

1. ***I'll give the stock recitation***
2. ***then, my own twisted paranoia***

first, our data formats are tiered

This really drives the analysis strategy.

Raw data – 250KB/evt - on tape at Fermilab

DST – 150KB/evt - on tape at Fermilab

- can be source of specialized ROOT tuples for physics
- can be source for hopefully most reprocessing

  for example, track clusters are saved, not raw hits

Thumbnail – 10KB/evt

- originally: "…rapid…event selection with rather minimal analysis". - for mining.
- But

  some groups analyze TMB files

  some groups unpack once into specialized formats, usually ROOT trees

Derived sets - ~few to 10'sKB/evt?

- this is still evolving from within the physics groups

Now for the history...

motivations have changed

During winter ‘01-’02, led by Jae Yu

- DØ*Remote*A*nalysis*C*oordination*E*ffort*
- code distribution, “empowering the remote desktop user”
- data distribution - setting up remote SAM stations
- concerns about network capability to FNAL
- more CPU - event remote primary reconstruction
- more CPU - remote reprocessing
- have R*egional*A*nalysis*C*enters* serve locally the role that DØ FNAL serves locally

Evolved through a RAC subgroup - current Task Force

- data distribution - tiers were finally fleshed out
- adding CPU - reprocessing from DST’s. Requires:
  - DST’s available on disk
  - database distributed
- Thumbnails on disk, distributed
- adding CPU - derived dataset production

RAC original idea was memorialized last summer:

DØ Note: 3984

Proposal for DØ Regional Analysis Centers

I. Bertram, R. Brock, F. Filthaut, L. Lueking, P. Mattig,
M. Narain , P. Lebrun, B. Thooris , J. Yu, C. Zeitnitz

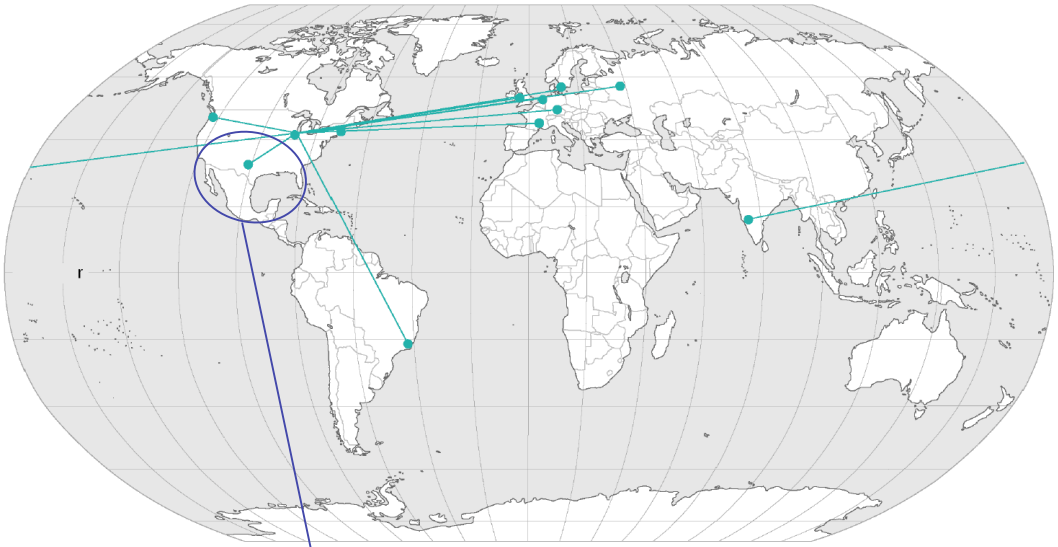
July 5, 2002

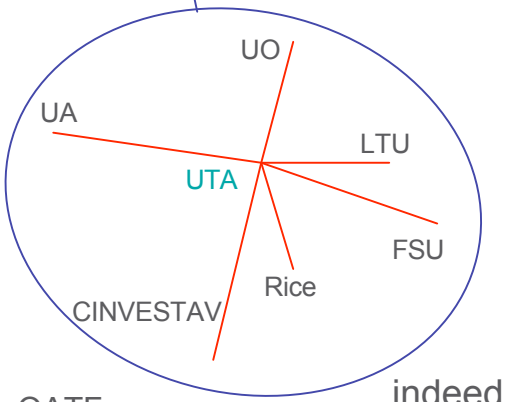
Abstract

The analysis of data from Run II will be such a significant effort, that in all likelihood it cannot be done by relying on the Fermilab site-based systems alone. Rather, success in

the basic scheme:

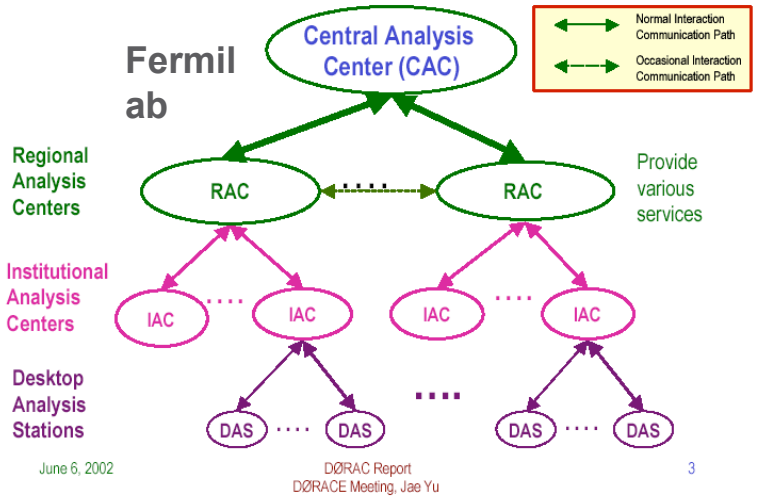
Initially, this (~10 RAC's?):



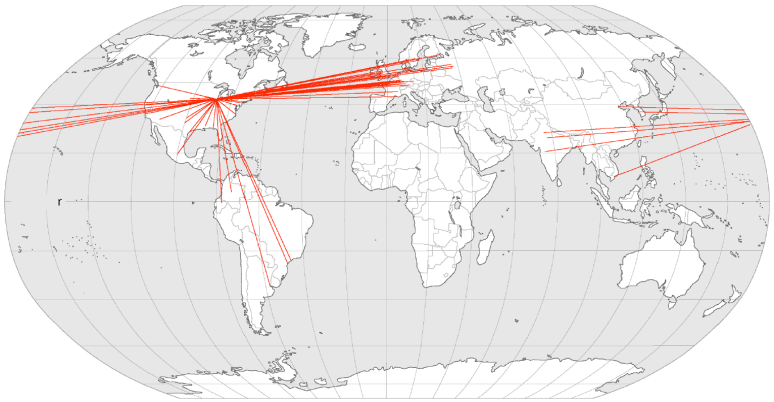


(arbitrary imagined distribution)

indeed, UTA has resources:



Rather than this:



minimum requirements (“D” class RAC): *computing/storage*

data storage - supporting analysis, first in its region

1. **all DSTs on disk at the sum of all RAC’s** – distributed randomly
   - qualitatively different from FNAL, rather complimentary to FNAL
   - hopefully the source for most reprocessing needs
2. **all TMB files on disk at all RACs, to support mining needs of the region**
3. **other formats on disk, according to the needs of each region**
   - derived formats
   - MC caching
   - database & SAM cache
   - temporary working cache ~10% of total

results in ~30TB disk storage/year per minimal RAC for Run IIa

cpu capability - supporting analysis, first in its region

Guessed 5% x recent DØ planning model

approximately 25 Linux nodes/year per D RAC for Run IIa

D RAC data storage

a model for storage:

	size	tape factor	disk factor
raw event	0.25 MB	0	0
raw/RECO	0.5 MB	0.001	0.005
data DST	0.15 MB	0.1	0.1
data TMB	0.01 MB	1	2
data root/derived	0.01 MB	0	1
MC D0Gstar	0.7 MB	0	0
MC D0Sim	0.3 MB	0	0
MC DST	0.3 MB	0.025	0.05
MC TMB	0.02 MB	0	0
PMCS MC	0.02 MB	0	0
MC rootuple	0.02 MB	0.3	0.1

multiples, or fractions of the raw event count in various formats

for example, this means:

1 complete data set-worth of TMB on tape;

2 complete data set -worth of TMB on disk

obviously, this is tunable

D RAC storage, cont

Disk Storage

	1 day	1 year	phase 1 2 years	phase 2 4 years
event rate	2.16E+06	7.88E+08	1.58E+09	6.31E+09
TIER DISK data accumulation (TB)				
raw event	0.0000	0.000	0.00	0.00
raw/reprocessing	0.0054	1.971	3.94	19.71
data DST	0.0324	11.826	23.65	118.26
data TMB	0.0432	15.768	31.54	157.68
data root/derived	0.0216	7.884	15.77	78.84
MC D0Gstar	0.0000	0.000	0.00	0.00
MC D0Sim	0.0000	0.000	0.00	0.00
MC DST	0.0324	11.826	23.65	118.26
MC TMB	0.0000	0.000	0.00	0.00
PMCS MC	0.0000	0.000	0.00	0.00
MC rootuple	0.0043	1.577	3.15	15.77
cache	0.0139	5.085	10.17	50.85
db/SAM		0.500	1.00	2.00
total storage (TB)	**0.1393**	**50.852**	**102**	**509**
total storage (PB)	**0.000**	**0.051**	**0.10**	**0.51**
total storage (GB)	**139**	**50,852**	**101,704**	**508,518**

Run IIa

Run IIb

Tape Storage

	1 day	1 year	phase 1 2 years	phase 2 4 years
event rate	2.16E+06	7.88E+08	1.58E+09	6.31E+09
TAPE data accumulation (TB)				
raw event	0.5400	0.000	0.00	0.00
raw/reprocessing	0.0011	0.394	0.79	3.94
data DST	0.0324	11.826	23.65	118.26
data TMB	0.0216	7.884	15.77	78.84
data root/derived	0.0000	0.000	0.00	0.00
MC D0Gstar	0.0000	0.000	0.00	0.00
MC D0Sim	0.0000	0.000	0.00	0.00
MC DST	0.0162	5.913	11.83	59.13
MC TMB	0.0000	0.000	0.00	0.00
PMCS MC	0.0000	0.000	0.00	0.00
MC rootuple	0.0130	4.730	9.46	47.30
total storage (TB)	**0.6242**	**30.748**	**61**	**307**
total storage (PB)	**0.001**	**0.03**	**0.06**	**0.31**
total storage (GB)	**624**	**30,748**	**61,495**	**307,476**

the cpb model presumes:

25Hz rate to tape, Run IIa

50Hz rate to tape, Run IIb

events 25% larger, Run IIb

minimum requirements ("D" class RAC): *databases*

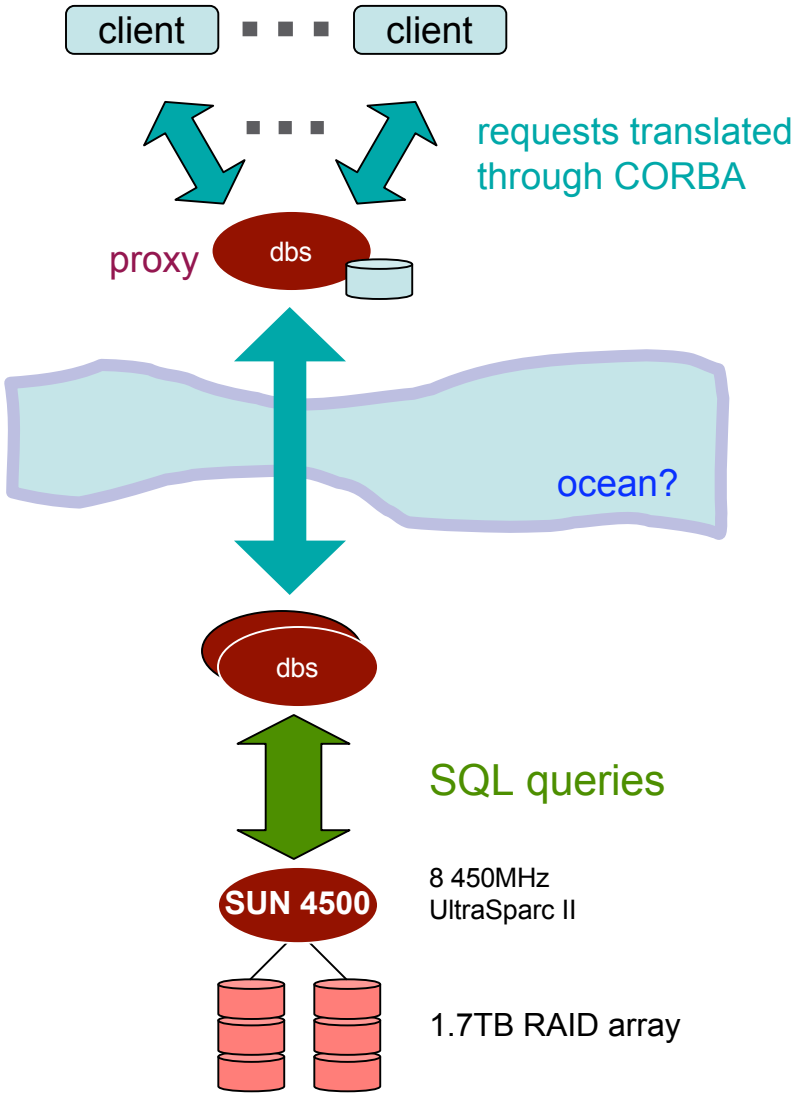
Oracle database access

Reply on proxied database servers

- a feature of the current server upgrade

  every RAC would house a proxy server

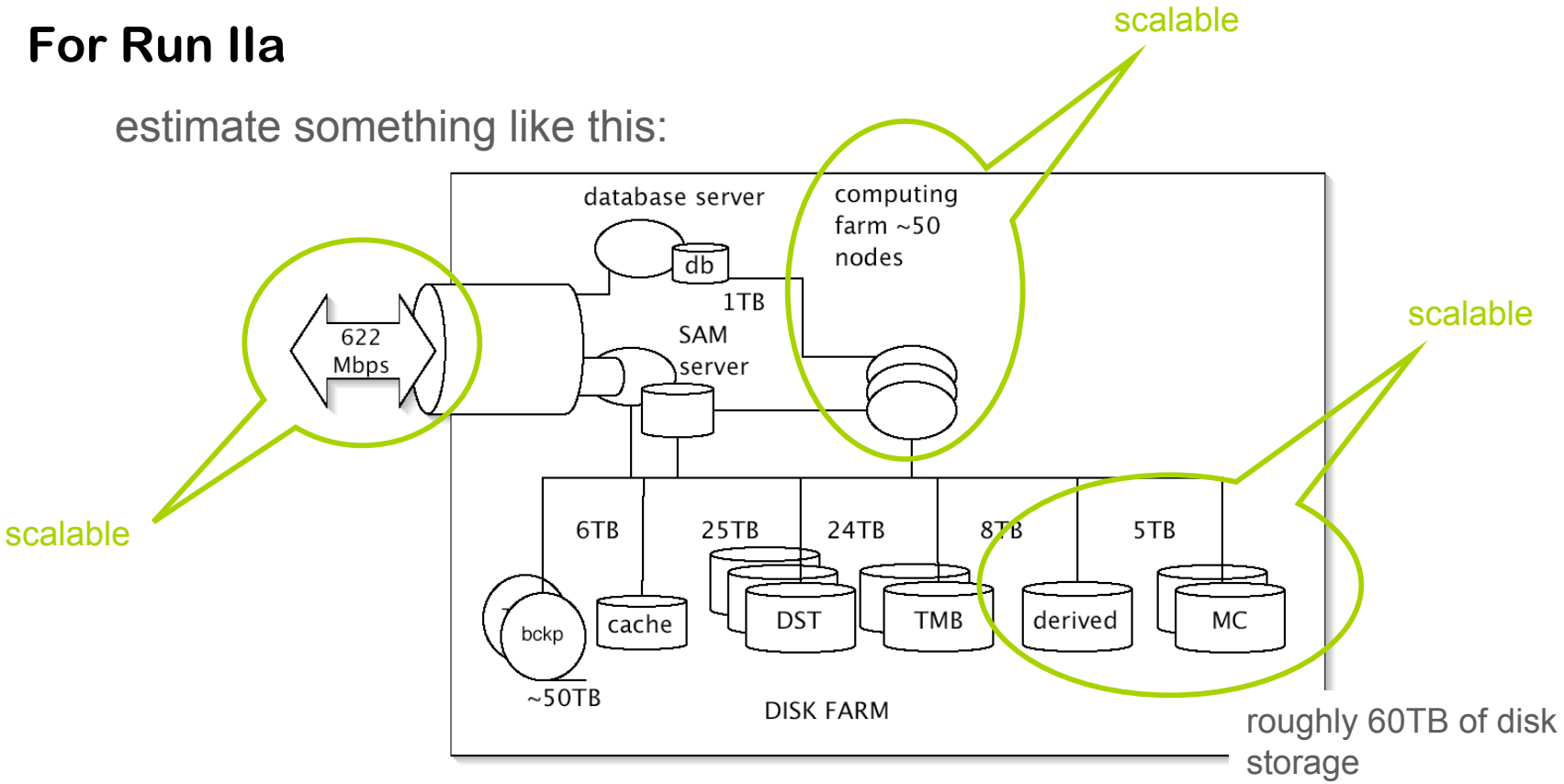
- this is being tested right now.



1 + 2 summary of the *minimum* ("D" class) RAC

For Run IIa

estimate something like this:



- this alone adds > 500 cpu's, deployed in an efficient way - where the physicists are
- IAC's should have have additional, considerable capability
- All in host countries.

scale it up…to the Best RAC ("B" class)

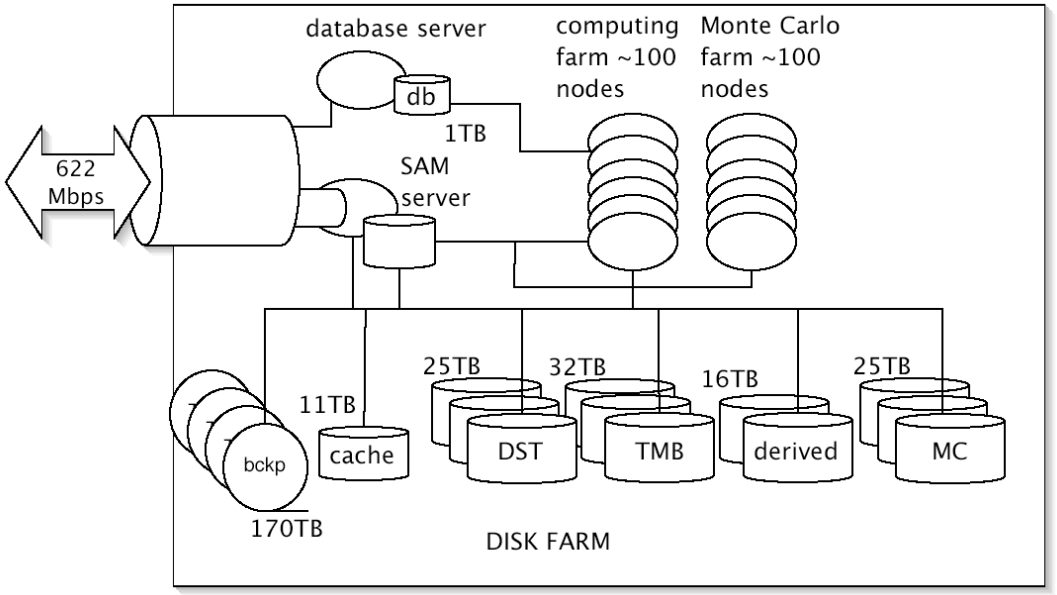
Keep: DST storage as a common resource, SAM, db proxy server

Add:

- More batch computing: +50 nodes
- MC generation: + ~ 100 nodes
- More MC storage: +20TB
- All TMB: +8TB
- More derived data cache: +8TB
- More temporary cache: +5TB

So:

~100TB of disk and ~200 cpu's



a very serious system

desirable to have a few

may fit as parts of larger facilities

- No longer manageable by a single university department

Task Force

...has thought about this for 8 months

- actually, thought a lot about the financial arrangements among institutions

...essentially endorsed all of the recommendations of the RAC document

...will be delivered to Spokesmen, or has just been.

I think that there are still missing issues.

I have a slightly skewed view:

I keep having this nagging feeling that our *real* problems are not:

- how to install more CPU
- how to install more disk space
- how to distribute the databases

That is, we're not SETI

- our problems will not be solved by just scaling our hardware deployment

Our Real Problems are much harder. They include:

- how to understand the underlying event consistent with 5 MeV/c^2 statistical precision on M_W
- how to find another factor of two reduction in Δm_{top} jet scale
- how to tag and vertex *B* mesons in an environment of 5-10 overlapping events
- and other similarly terrifying obligations

These are problems that require people working together in a different way from the past

Infinite hardware - Grid - resources will not solve these real problems.

Just as the best politics is local,

physics analysis is an intense, interactive process

- projects can factorize sometimes, but it is delicate and requires lots of communication
- without communication - maybe even communication tools - the hardware resources will be inefficiently used

So, the goal

- to support groups who are within efficient people-in-contact distance

  by:

  engaging the senior collaborators in real time

  - *senior European colleagues can now either go nowhere, or 2500miles.*
  - *if RAC works, they will also have opportunity to be productive traveling a few hundred miles to be with other people*
  - *RAC's can be a point of contact*

  engaging the students before they leave for Illinois and after they return

  empowering the extended collaboration by making the Regions capable

centralized analysis? *I'd answer, 'hope not...'*

The full exploitation of multiple fb^{-1} of luminosity

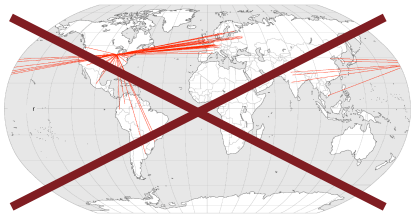
- forces us to confront enormous challenges of

  - heroic calibrations, terrifying systematic uncertainty requirements, etc

- which will require **many** people working **closely** together

  many ⇒ more than the ~200 seats at Fermilab, *ie. more than just the Run I commitment*

  closely ⇒ more intensively than would be possible in one, big FNAL-centric cluster , ie not:



Said slightly differently:

Our problems are hard.

Solving them will require intense, personal collaboration.

So, the RAC group tried, but I think failed

to figure out how to bring FNAL functionality to the whole DØ world

and make it usable in the invent-it-as-we-go environment of analysis

- none of us thought about this at the beginning..

first things first:

We still need to think about what's required to do our analysis, socially (if you will), as well as technically

…and plan to that.

Our concept has evolved in parallel as our computing model(s) has evolved - driven by what's really going on in the Physics groups

- I think that these evolutions are not in synch yet.

Our distributed computing architecture

should follow a realistic analysis strategy.

That is, our analysis strategy should not inherit an architecture.

I would like to see us concentrate on more than just aquiring and linking *stuff:* how we will use *stuff*

- with a system for <u>people that grows</u>
- accommodating our learning curve, and paced to our opportunities.
- merging to a true grid as it becomes available
- This implies an organization which takes the physics as primary.

the questions are:

The *Hustlers* question:

How do we do that?

The *Longest Summer* question:

How do we stage it?

The *Tin Cup* question:

How do we pay for it?

The *Field of Dreams* question:

If we did it…would remote institutions come? That is, would they scale their active, intellectual engagement?

The *When We Were Kings* question:

How should the structure of DØ's physics management change?

The *Raging Bull* question?

Actually, I couldn't think of one, but wanted to get Jae Yu's name in here somewhere...

So, you say, isn't this the Grid?

Yes: If you think like me that a functioning Grid is people, AND hardware.

- So, yeah, we say "Sure, it's the Grid."

But, in practice, we have to be careful.

- The full Grid - *opportunistic computing, transparently brokered over the globe, with full access to data, code, and databases* - is going to happen

  But, our timescale is now and our use of the Grid will have to be learned.

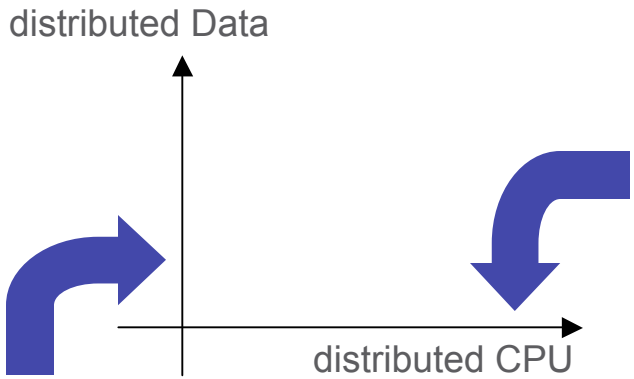
- Rather, if we can stage our way toward implementation of Grid tools and capabilities, we can
  - move deliberately and learn to work globally, by starting regionally
  - pick and choose what and where Grid tools are tested/used
  - **...control our risk while maximizing our capabilities**

DØ analysis of Run II cannot be a beta-site for Grid prototypes.

I think that the R(egional)AC concept allows this...

aside: what's the Grid?

I'm learning to think of the Grid as a 3d space. Not just:

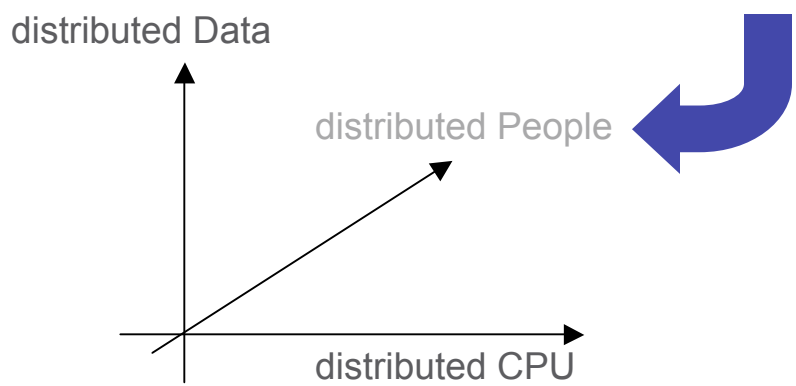


a lot of effort goes into learning to manage resources along this axis - this gets a lot of attention

SAM, provides us flexibility along this axis, and essentially does it now

but:

I think that this is new, at the scale of even DØ, let alone LHC



For us, because of our backgrounds and strengths, a deliberate Grid deployment takes us out the

a) Data axis first (we know that dimension pretty well), then the

b) People axis next (the hardest thing, but we'll learn), and finally,

c) the CPU axis, as we need it

getting there deliberately

The core essentials and time-ordering:

1 minimal requirement: initially defines an RAC

- the **data**

  in all of the formats that we anticipate necessary for mining, analysis, calibrations, and derived data set production
- the **databases**
- **sufficient CPU** capability to mine/select/manage the data
- the Regional **network**

minimal requirement: soon specifies a system of RAC's 2

- the Global, RAC-RAC **network**

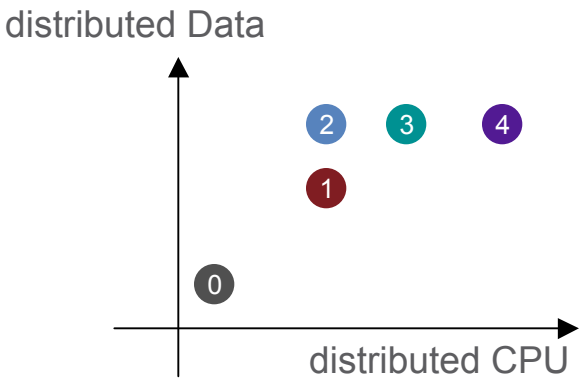
ultimate requirement: characterizes a REGION 3

- **enough CPU** capability to reprocess from DSTs at least

- **IAC CPU** capability for the entire collaboration, irrespective of Regional home

4 ultimate requirement: completes the DØ VO

I'm suggesting:



Remember:

0: now

1: defined RAC

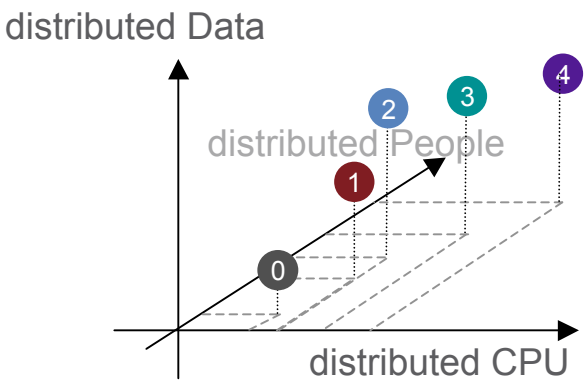
2: RAC-RAC networking

3: reprocessing power

4: full REGIONs, increasingly Gridified, *eventually* Remote Analysis Centers

The real need for CPU will come in time, so…do the hard part first:

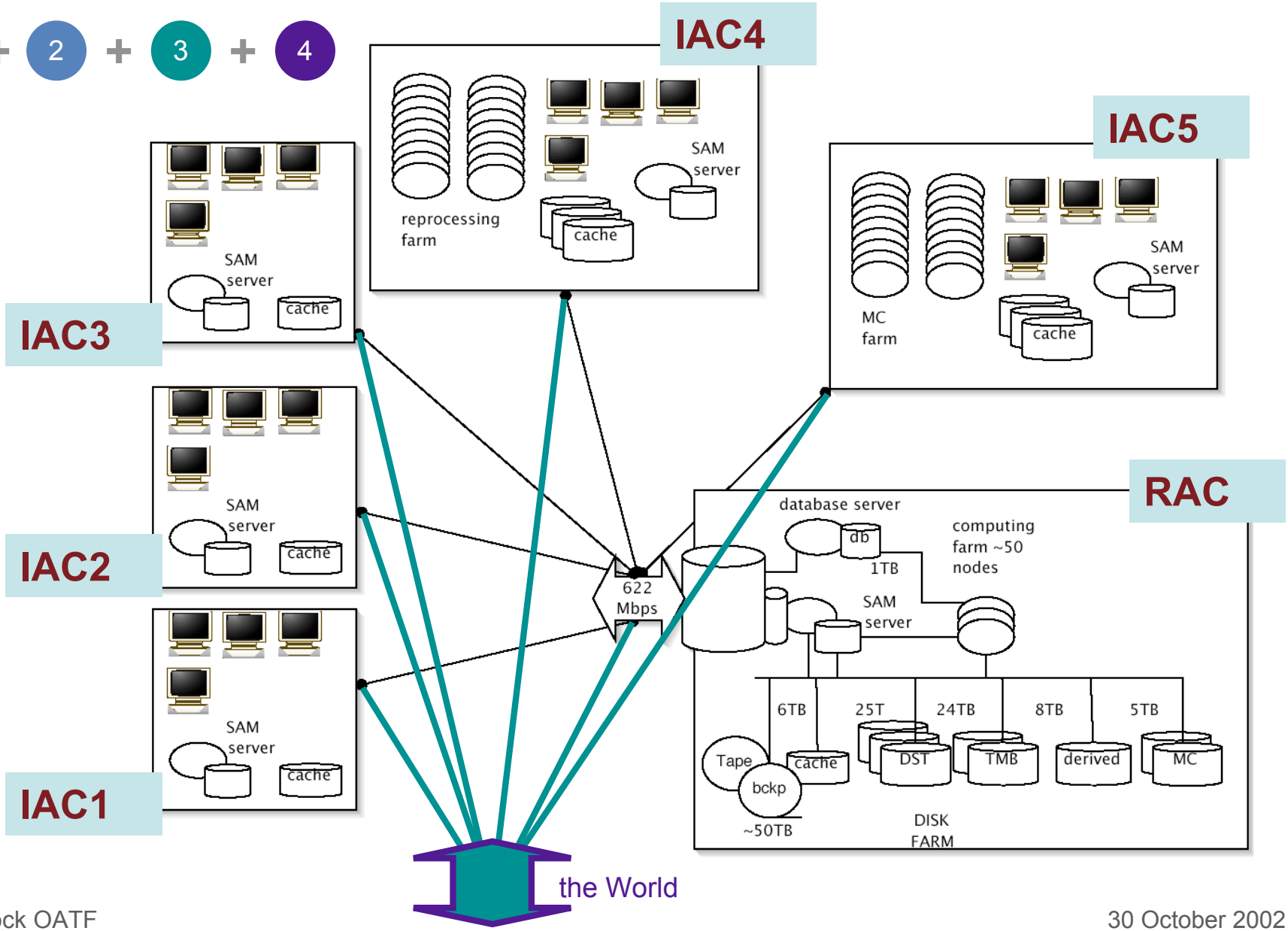
- *Increase along the People axis*
- *develop the new paradigm for organizing the analysis*
- *accommodate the inevitable organizational complexity*
- *investigate the whole area of* ***Collaboratories***



it's the *n* REGIONs that count

first: then:

1 + 2 + 3 + 4



summary

I'm suggesting a strategy of replicating FNAL analysis intensity around the world.

That to do so requires staging in order to:

1) treat Job One as that of organizing people and what it means to do physics analysis–to learn to work along the third axis;

   *I think that initial organization of remote analysis **within Regions** does this best, with the most long-lasting benefits*

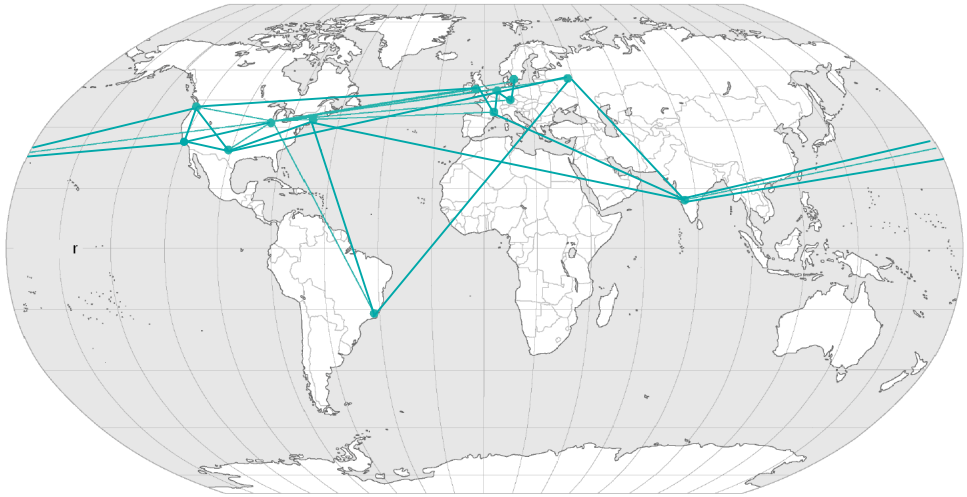
2) aggressively participate in Grid development, incrementally deploying tools (scheduling, brokering, collaboratories) as we need them and as we can handle them; and

3) finally, as we learn to do analysis remotely, indeed seek to be the Grid:

   ***evolve REGIONAL** ANALYSIS CENTERS **into REMOTE** ANALYSIS CENTERS:*

(arbitrary imagined distribution)

BUT



Brock OATF