

Offsite Analysis for Run II

...making the Regional Analysis Center concept work

plan for this discussion

1. general motherhood statements
2. specifics of the RACE group's characterization of:
 - services to be provided by a RAC
 - capabilities that a RAC should have
3. concerns that I've got
4. conclusions

so,what is it?

Regional Analysis Center (**RAC**)

is an off-site facility that serves as a hub to nearby

Institutional Analysis Centers (**IACs**)

there is a document:

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

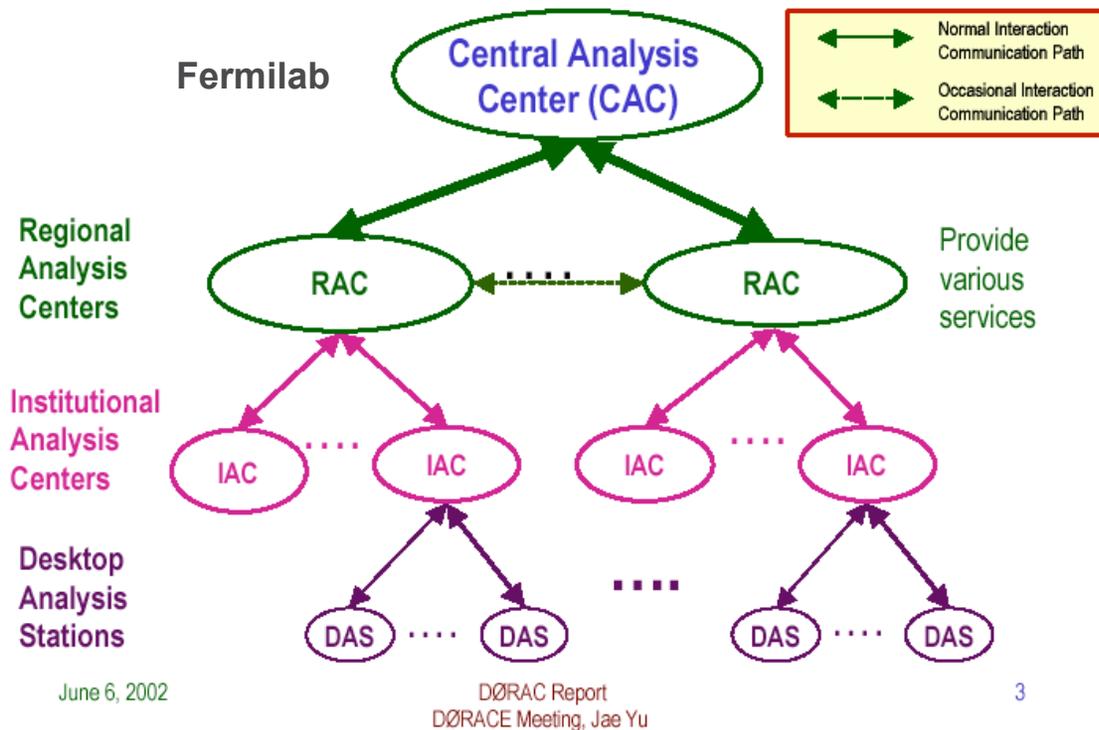
This discussion follows that paper and the consensus that led to it

so, what's the RAC universe?

Jae's model

regional clusters of analysis institutions served by central, capable center...Regional Analysis Center (RAC)

Proposed DØRAM Architecture



I can get to 10-11:

- one each in Great Britain, Germany, France, Russia, and the Netherlands
- one in South America
- one in Asia (including India)
- one at Fermilab (CluDØ/CluB?)
- 3 more in the US - East, West, South

presumably distributed according to geographical, political, and/or infrastructure criteria...

so...why?

Because of the promise and complexity of the Run II data

The worldwide investment in this run demands that we get all of the physics out

The old way...where the action's @FNAL:

- almost everything done at FNAL, outside institutions station as many people in Illinois as affordable, and the faculty travel - gotta be here.
- Fermilab absorbed the cost of processing and data storage - we got by

The RAC way...where the action's @everywhere:

- Off-site institutions become full intellectual contributors: **analysis is better**
 - presumably critical for off-shore groups, maybe even desirable for US groups?
- The physics analysis effort probably demands it
 - enormous luminosities place extraordinary demands on the analysis, which likely cannot be met by Fermilab alone - not in computing power, storage, or available seats
 - not just a question of just getting the answer: the systematic uncertainties in Run II must be consistent with the scary statistical precision*
- My opinion: the health of HEP on US campuses needs an @home-presence

so, is the argument air tight?

I think it could still use work...

Need Use Cases...which have the dual impact of:

- emphasizing how cool it could be and exposing the complications
 - The document contains a narrative for a W cross section measurement

Also need:

- a tracking Use Case, B lifetime .
- a high statistics Use Case, High E_T jets.
- a reprocessing Use Case.

A we'll-fail-without-it argument

- deconstruct a few Run I analyses - what was really done - and project them onto multiple fb^{-1}
 - started to extrapolate the Run I M_W analysis
 - I think that the amount of work required for all anticipated analyses will be impossible the Old Way

so, what about... **THE GRID?**

I know what you're thinking: "Is this the famous GRID?"

For DØ, probably not in its full glory – in our analysis lifetime

- can the enthusiasm of worldwide GRID proponents be justified? we'll see.
- **BUT** - some increasingly capable toolkit for resource balancing, job submission, data transfer, scheduling, statistics, metadata access, etc. *will come*, incrementally

We'll always be somewhere between no Grid and full Grid

- our experiences will **almost certainly productively feed back into LHC GRID planning and maybe global GRID planning**

We have a distributed data management tool now: SAM

- "GRIDifying SAM" or "SAMifying the GRID" is a major priority

With or w/out GRID, coordinating humans will be key

Need flexibility, replace humans with tools when stable and useful

Premium on stability, the analysis is not a GRID beta test facility

so, what are the imagined RAC services?

- **enhanced batch processing for region**
maybe IAC processing privileges at local RAC initially?
- **data cache and delivery for region**
RACs deliver not just to local IACs, but everywhere
- **database access for region**
hopefully can rely on db proxies
- **data reprocessing for collaboration**
- **monte carlo production, or service to related MC IAC sites**

notice what's not there:

ab initio reconstruction

- presume farm will always keep up

code distribution

- after discussing it, there seemed to be no necessity for RACs to support code distribution outside of the currently evolving UPS/UPD based distribution started with the DØRACE workshop
- there might be a need for local support structures to triage questions/problems before they get back to FNAL
 - presumably distributed expertise with code dist., SAM, databases, etc.

so, what constitutes an RAC site?

use the Sears model:

“good” - *some minimum capability, to be determined*

“better”

“best”

- I'll characterize “**Best**”...then imagine a continuous scaling to “good”

The bottom line for the system of RACs:

the totality of RACs would have to be capable of:

- reprocessing the data if required
- complementing, not just replicating the FNAL storage capacity
- significantly increasing the intellectual input to the whole analysis

Best RAC requirements, 1

location, location, location

They have to be positioned in order to serve

- Anticipate a few RACs - not more than ~10

Didn't try to establish firm siting criteria

- rather, try to distribute according to density of users
- there will be other overriding considerations:

network capabilities, political issues (language, funding, national goals, etc), physicist interest, etc.

Networking capabilities

high-bandwidth, RACs to FNAL required

high-bandwidth, RAC to local IACs

nice, but not necessary, high-bandwidth, RAC to all other RACs

Best RAC requirements, 2

data storage

Generally thought desirable:

- all TMB files on disk at all RACs
- all DSTs on disk at the sum of all RACs -distributed randomly
 - *qualitatively different from FNAL service - complimentary*
 - *hopefully the source for most reprocessing needs*
- a variety of other formats on disk, keeping in mind MC needs may involve local, high-capacity caching
 - *rootuples or other derived formats*
 - *MC DST – depending on MC generation within cluster?*
 - *database/SAM disk storage*
 - *temporary cache ~10% of total*

results in ~50TB disk storage per year per Best RAC for Run IIa

computing. Used cpb model, guess \square 10% x fnal capability

guess ~50 nodes per year per Best RAC for Run IIa

Best RAC data storage

using the tools for the cpb document to the Director's Review - 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

for example, this means:

← 1 complete data set-worth of TMB on tape;
2 complete data set -worth of TMB on disk

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

obviously, this is tunable

Best 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

Best RAC requirements, 3

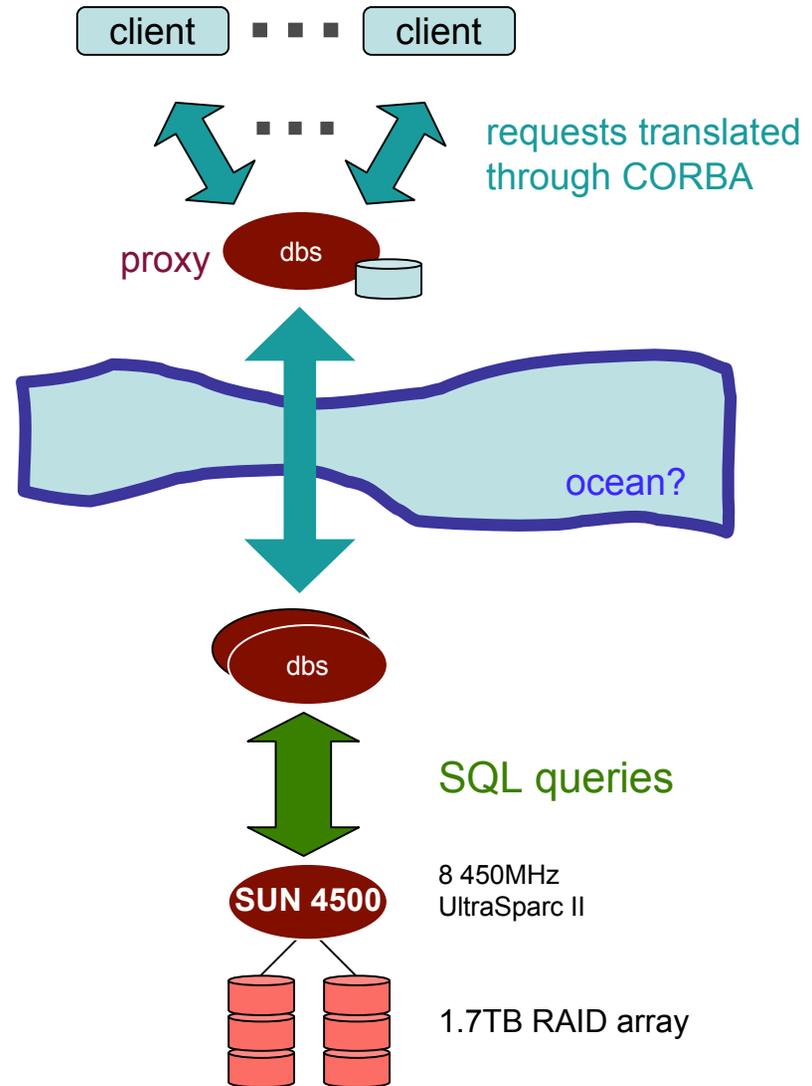
database

presume implementation of proxied database servers

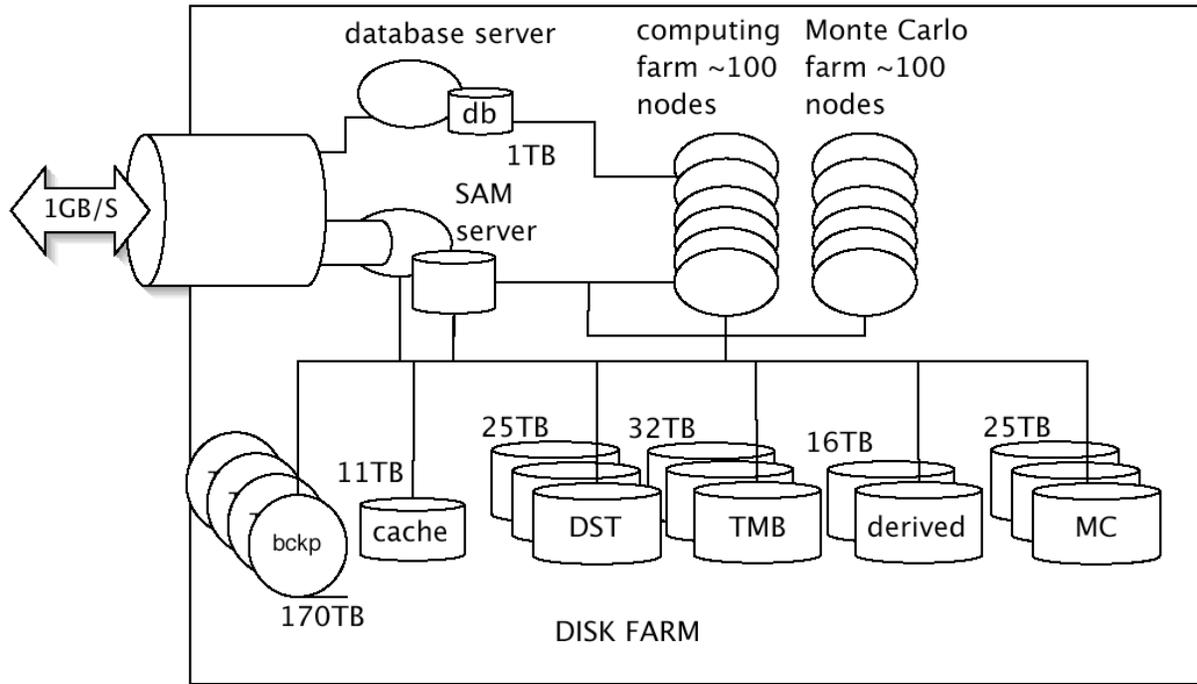
- a feature of the upgrade currently under way for the server

every RAC would house a proxy server

- this will hopefully be tested within the year



so, what's the summary of the *Best RAC*?



probably ~\$1M center...sobering, especially given support required

BUT...jeese: the GRID/LHC business is a very expensive, and well-supported enterprise - surely this fits in that planning?

There has to be an argument that investing in a real experiment will guarantee that the LHC GRID effort will be better.

This might be worth thinking about among joint DØ-Atlas/CMS institutes.

so, the other Sears categories?

Good:

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

Reduce:

- Much less MC storage: – ~20TB
- Maybe not All TMB, but particular streams of TMB: – ~8TB
- Less derived data cache: – ~8TB
- Less temporary cache: – ~5TB
- Less batch computing: – ~ 50 nodes
- No MC generation: – ~ 100 nodes

So: ~60TB of disk and ~50 processors ...~\$300k?

- Manageable by a single university department?
- Does this make sense as a minimal system?

Hey, maybe we can have you in a like-new, fully equipped RAC for \$300k-\$1000k

reprocessing: is this a requirement? *most think that it must be*

Could be a major headache, or worse

- It happened a few times in Run I
- FNAL reconstruction farm will be fully busy with data coming in
- What would we do if we discovered a problem that required reprocessing from raw data?
 - study it hard...certainly no decisions over night
 - could decide to spend \$ and triple the size of the farm and just do it *in situ*
 - or, could decide to use the set of RACs
 - involves getting raw data to them*
 - o at that time? Major issue of organization, heroic 24/7 robotic gymnastics, and significant ethernet traffic. Maybe doable. Painful.
 - o plan for it and continuously ship raw data to all RACs? Would require 400TB ÷ #RAC ~ 40-50TB of tape storage...not trivial

makes the design of the DST really an important exercise

- reprocessing from DST would then be relatively straightforward at RAC's

so, we got still more requirements

Need support - *Best* would require serious professional help

Not a simple setups - requires committed system management

Sharing with other experiments (not just DØ...not just HEP...or not even physics!) - is inevitable.

this probably has its good aspects and its not so good aspects

- loan of resources if crisis? funding? collaborative GRID R&D?
- But competing for resources and living with the politics of the GRID biz might be frustrating.

Need a serious MOU structure

1 RAC dropping the ball freezes out the IACs and affects all of DØ

Need a worldwide management structure to keep the whole thing moving toward results.

– Will keep spokespeople & physics coordinators awake at night...

so, we had a suggestion:

We thought of commissioning a *Prototype RAC Project*

– Identify, hopefully, a European institute (RAC₁) and a set of committed regional institutes (proto-IACs)

– Three goals:

- TMBs are shipped in real time, continuously
- and used by the proto-IACs to do physics
- do it by winter?

– Declare success

- Autopsy the effort and do it better the next time

I suspect: sociology and management will present as big a set of complications as technology

we need to understand this

volunteers?

so, what are my concerns?

- **Is this a problem that needs solving?**

A serious discussion has to happen before embarking

- **Can you say “Video Conferencing”?**

This is being worked on, but will cost money.

- **Culture**

The collaboration and the Community need to buy into the idea that it's okay to be a post doc or student and live off-site.

- **At the risk of repeating myself:**

The **management** of a world-wide analysis done in this fashion would be unlike anything HEP has tried to do before

- However, when it comes to the LHC, we talk calmly like “no problem” for 1500 close friends to collaboratively analyze experiments. Let's test that.

- **John and Jerry need to be sure about this!!**

okay, so I worry a lot

- **Review the few international computing projects that we have going now:**
 - From both ends...how have they gone?
- **How do we Decide?**
 - Um, I mean, it's not a voting matter... so what do we do?
- **The US funding agencies need to recognize it**
 - LHC needs to embrace such an effort as strengthening their pie-in-the-sky plans for this sort of thing (editorial)
- **The draft document currently has 28 conclusions that should be considered**

decisions required: 0. - 5.

12.1. Summary of Conclusions

Conclusion 0. Remote analysis capability with full access to the data, code, and collaborative analysis is necessary in order to satisfy the physics goals of Run IIa and IIb. A structured environment which systematizes and standardizes these services is the best way to implement this program.

Conclusion 1. It is anticipated that the FNAL processing farm will be sufficient for all of Run II primary reconstruction needs. RAC's are not envisioned for *ab initio* event reconstruction.

Conclusion 2. RAC-centered resource management is an important goal. While initially resource management may require considerable human organization, it is desirable to augment and replace that intervention with emergent GRID tools. The priorities assigned to tool deployment remains to be worked out with sufficient Use Case analyses and some real-world experience. Accordingly, the actual capabilities of the evolving system need to be carefully planned, biased toward smooth running rather than alpha or beta testing of GRID sites.

Conclusion 3. Continued evaluation of the number of off-site potential users and their anticipated needs should be undertaken very soon. A preliminary census has been done. The follow-up should include more detailed scenarios and/or capabilities for a more realistic assessment.

Conclusion 4. A complete review should be done of the planned data tiers with special attention paid to potential off-site reconstruction opportunities with DST's and analysis opportunities with a TMB's. This should be done before deploying the DST/TMB files.

Conclusion 5. Generally, RAC's need not be the sole sites of code distribution to their IAC's. Rather, at least for the early days, individual installation and updating can be done directly from the Fermilab site.

decisions required: 6. - 11.

Conclusion 6. Robust versioning and a scheme for guiding or automatically initiating stale file and directory deletions should be designed as soon as possible.

Conclusion 7. It may be important to precisely assess the degree of database access required for Monte Carlo production capability which includes overlaid events.

Conclusion 8. Early generation tools for interrogating the MC farm sites for available capability and tools for referring the actual job submission to those waiting sites are required now. Evolution of this system into a single step with full GRID deployment should be envisioned when it becomes available.

Conclusion 9. MC generation at RAC and/or IAC sites will be necessary in an increasingly wider scale as systematic uncertainties become a focus of measurements. Sufficient bookkeeping capabilities with the flexibility to imbed and modify the MC generation details will reduce the re-generation of already existing scenarios.

Conclusion 10. For the meantime, sufficient batch processing capability should be at each RAC to serve the needs of only the regional IAC's. What is uncertain is the amount of processing that this will entail and an effort to quantify this should be undertaken. Estimates will be made below.

Conclusion 11. In general, it seems reasonable to presume that in view of the limited computing resources available at FNAL, and of the improvements that are likely to be made to the algorithms used at present for the reconstruction of the collider data, some measure of reprocessing is expected to be an essential ingredient of the RAC's.

decisions required: 12. - 15.

Conclusion 12. Answers to questions 1-7 need to be in hand before reaching a conclusion about how to fully characterize generic RAC's.

1. How can the DST be ensured of serving as a useful basis for reprocessing? This is a timely design issue.
2. How will DST's be distributed for this reprocessing?
3. What strategies for raw data-level reprocessing can be designed?
4. If such strategies require RAC participation, how will the raw data make their way to the RAC's?
5. For any reprocessing operation, for the first time original data will reside at a location (many locations!) away from Fermilab. Do these derived data sets get transferred back to the central Fermilab facility? Certainly, the answer is "yes" for the TMB, by design.
6. Obviously, such a scheme involves significant networking and bookkeeping resources at the RAC's and the affordability of this should be understood.
7. How can the consistency of the reprocessing be guaranteed? This is not a matter only of code distribution, but also of hardware, OS, etc.

Conclusion 13. Reasonable database services could be achieved with the development of proxy database servers at each RAC.

Conclusion 14. Full testing of the performance of the new DBS implementation should be performed at the soonest available time.

Conclusion 15. A test installation of the proxy server idea at a remote site should be done in the near term.

decisions required: 16. - 22.

Conclusion 16. An evaluation of networking needs for remote analysis should be done for FNAL, U.S. RAC's, and overseas sites.

Conclusion 17. A robust and reliable SAM system at every worldwide site is essential. This means that adequate support for both development and operations must be provided.

Conclusion 18. All TMB records should be disk resident at all RAC sites twice. Total for Run IIa for TMB storage of 16TB disk per RAC.

Conclusion 19. Significant storage for project formats should be available at RAC's of the same order as the total TMB cache. Total for Run IIa for DERIVED storage of 16TB disk per RAC.

Conclusion 20. Complete DST data formats should be disk resident within the sum of the RAC sites: Total for Run IIa for DST data storage of 24TB disk per RAC, which presumes 10% of the total at each RAC site.

Conclusion 21. MC storage for per-demand generated events is primarily limited to the ROOTuple needs, with a nominal MCDST compliment as well: Total for Run IIa MC ROOTuple data storage of ~5TB disk per RAC and 10TB tape per RAC; MCDST disk storage of approximately 50TB, presumes 5% of the dataset on disk.

Conclusion 22. Temporary storage needs for staging of data and Monte Carlo analysis and ROOTuple generation are estimated to be 10% of the total of each data format: Total for Run IIa for temporary cache of ~11TB disk per RAC. A more accurate estimation of this need is required.

decisions required: 23. - 27.

Conclusion 23. Storage needs for serving db setups at RAC will not likely exceed a few TB for all data taking. Total for Run IIa for database/SAM needs of 1TB.

Conclusion 24. Batch resources per Category B RAC should be on the order of 50 nodes per year of modern PC nodes running Linux.

Conclusion 25. Planning should begin early for constructing and maintaining a sophisticated electronic helpdesk and FAQ for DØ software installation, implementation, and use issues. Triage strategies should be planned.

Conclusion 26. It might be useful to get a non-binding expression of interest from potential institutions just to see what the maximum might be, and to determine whether interest is not sufficient to support the concept. Too few sites will make the burden on that small set perhaps too significant for their viability.

Conclusion 27. A **Prototype RAC Project** should be mounted to establish a working RAC by October 1, 2002. "Working" should be minimally defined to be a) the RAC site accepting continuous TMB files; b) and identified IAC sites in a new cluster using them with relative ease to do DØ analysis.

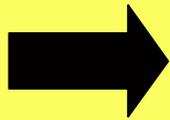
continuing to worry...

- *Did I mention management?*
- **How do we convincingly assess the interest overseas?**
 - My suggestion is that this is a big enough deal that:
 - some number of spokesmen go overseas *soon* to all off-shore DØ countries and ask for help in the Run II analysis along the RAC plan and
 - offer the assurance that this partnership will have the full backing of the experiment and the Laboratory

conclusion

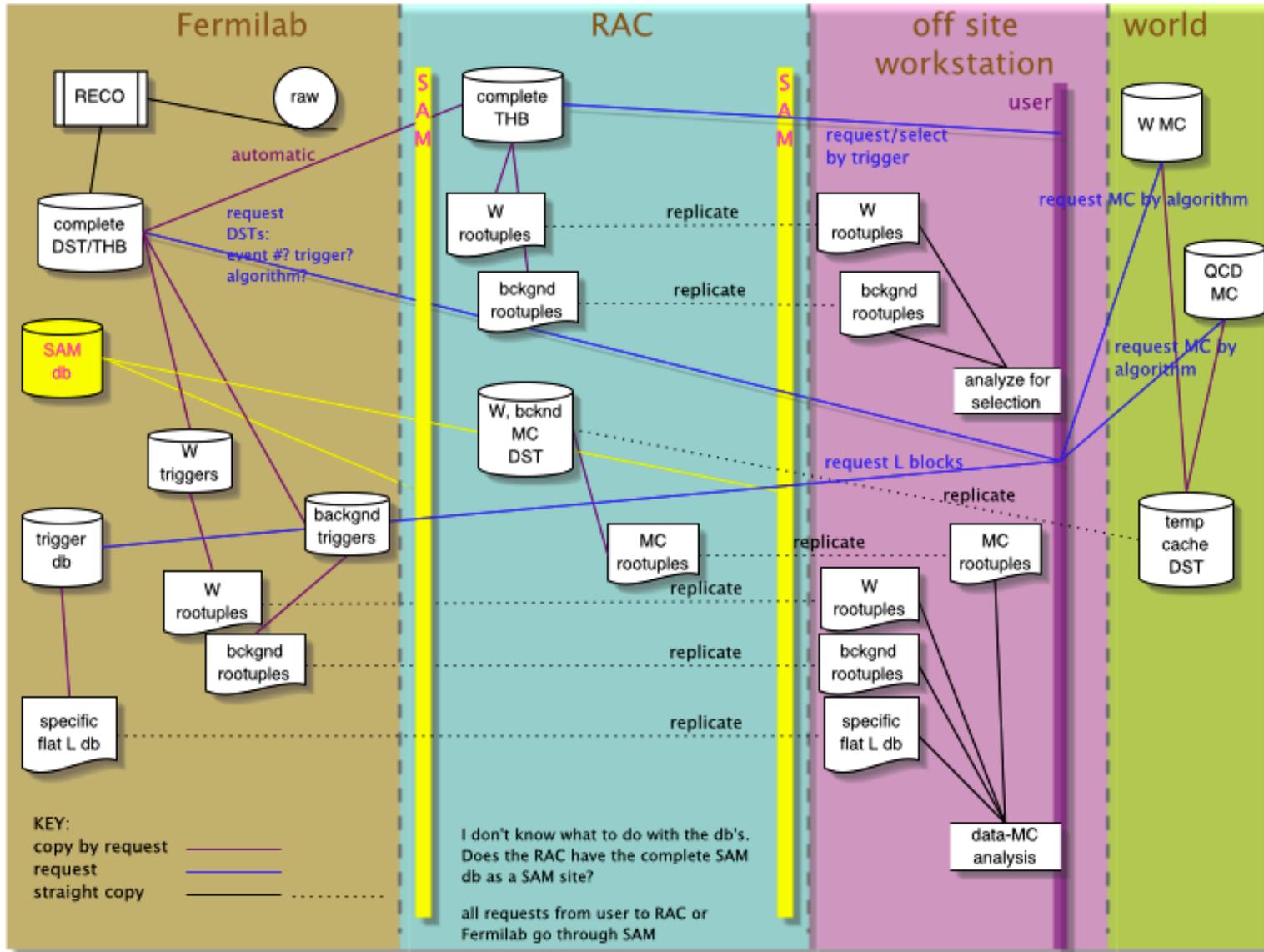
This is early days -

- but the idea was floated with the Run II Computing review:
 - “Both collaborations [should] develop more detailed plans for the coordinated use of remote computing facilities...”
 - “The Committee congratulates the [sic] Dzero ... on its aggressive strategy to develop Regional Analysis Centers that would provide centralized regional access to data analysis resources. CDF’s effort has been more modest...”
- we need to decide whether to do it and to what extent
 - we need a assessment of interest
- we need to explore cooperation with LHC/NSF/DOE
- we need to hear from you if you’re interested



analysis of Run II will be very tough if it is fully collaborative, worldwide.
analysis of Run II may not succeed if *not* fully collaborative, worldwide.

building Use Case for possible IAC desktop W cross section measurement



[back](#)