18th June 1999

PROPOSAL FOR A CMS(UK) COMPUTER CENTRE

Greg Heath, Helen Heath, Dave Newbold & Glenn Patrick

1. Executive Summary

We present a first discussion of the computing needs of CMS(UK). The immediate motivation for this is that we are about to enter discussions with ATLAS and LHC-B on a possible LHC-wide bid for JIF funding for computing resources. We need first to reach an agreement within CMS(UK) on the overall direction of our computing strategy. Assuming that we can find some common ground with the other collaborations, we will then put together a proposal to JIF. Representatives of each of the three University groups will be asked to arrange for the necessary signatures to the proposal from their institutes.

Models for LHC computing are being discussed in the framework of the MONARC (Models of Networked Analysis at Regional Centres) group, from whom we take some of our quantitative input. These models consider the situation a year or so before and after LHC start-up. Bearing in mind the timescale for JIF funding, we have also tried to consider our needs in the more immediate future, and how we might build up toward our eventual goals.

Our analysis of the UK requirements is set out in detail in the following sections. Here we present a summary of our main assumptions and conclusions.

1. CMS computing as a whole is based around the model of "Regional Centres". In addition to facilities at CERN, there will be 5 or so major centres providing computing services to the collaboration. The UK does **not** have the resources to provide one of these centres.
2. Within the UK we will establish a **single** centre as the most efficient way to meet the needs of UK-based analysis teams. Even to provide such a "national centre" requires a major new development.
3. We have looked at a possible model for the setting up of such a centre, covering the period to 2007. Our rough estimate of the total capital expenditure required over this period is **£2M**. About 30% of this sum could be spent in the period up to 2003 and be the subject of a bid to the JIF.
4. In addition to this expenditure on equipment, the running of such a centre requires a number of dedicated staff. We estimate that **3 FTE** new staff posts will need to be supported for CMS.
5. The collaboration needs (or the three LHC collaborations need) to plan for the funding of such a facility. In order to make a realistic and successful bid to JIF for the initial part of the funding, plans for the required further developments, including both equipment and staff costs, must be in place.

2. The Regional Centre Model

In the MONARC project for LHC computing, several tiers of Regional Centre are defined, with a *Tier 1* Centre providing 20% of the CERN capacity for a single experiment and a *Tier 2* Centre providing a more restricted service. Informal discussions within CMS(UK) have so far rejected the idea of a Tier 1 Centre providing collaboration-wide resources, preferring instead to focus on the specific data analysis requirements for UK physicists. Moreover, it was felt that there would not be sufficient effort or resources to deploy a distributed computing model data throughout all four UK institutes. This has led us to the idea of a centralised CMS(UK) computing/analysis facility based at RAL.

We have made estimates of the requirements for data storage and computing power at such a facility in the years 2005–2007. Figures are available within MONARC for the likely size of computing facilities per LHC experiment at CERN. The size of a Tier 1 Centre is assumed to be around 20% of this central facility. Our estimates, discussed in more detail below, suggest that the required capacity of a UK Centre is roughly 5% of that available at CERN. We have also tried to estimate the requirements for 2003 and to look at how the facility might be built up starting from next year.

2.1 Physics Requirements

We assume that initial reconstruction of all data takes place at CERN, and that re-processing is shared between CERN and the Tier-1 Centres. The functions of a UK Centre after LHC start-up are to support UK-based analysis teams, and to contribute to Monte Carlo event production. In the more immediate future, the facilities would provide dedicated facilities for the simulation & optimisation of sub-detectors, physics studies and the production of OO code.

2.2 Usage of the Facility

Currently, the number of people in CMS(UK) performing large-scale computing is relatively small. By 2007, we assume that our facility is required to support the activities of 25-30 people working on analysis in the UK. Perhaps half this number might be active in 2003.

2.3 Obsolescence and Evolution

Most computing equipment will probably continue to have a working lifetime of less than 5 years. By continuously replacing obsolete equipment we should automatically achieve an increase in the CPU performance and disk capacity due to advances in technology. Replacing equipment as it ages will also allow us to cut down on maintenance costs. There has to be some mechanism for funding such upgrades if a UK centre is to remain a competitive entity.

2.4 Networking

For a centralised CMS(UK) facility to be successful, good networking between the university institutes and RAL will be essential. In addition, the transfer of hundreds of Terabytes of data each year will require excellent international networking with CERN or, at least, good infrastructure for the shipment and staging of transit tapes.

3. Technical Details

Here we develop our model for the storage and computing power required for CMS as a function of time. In the subsequent section we convert these estimates into rough costings.

3.1 Storage

To a large extent, the scale of a computing facility is dictated by the size of CMS event samples that need to be transferred from CERN and analysed in the UK. The table below gives the parameters for each category of event:

	Event Size	**Sample Size for 1 Year**
Raw Data	1 MB	1 PB
Event Summary Data (ESD)	100 kB	100 TB
Analysis Object Data (AOD)	10 kB	10 TB
Event Tag Data (ETD)	0.2 kB	0.2 TB

We have made the following assumptions:

- The raw data will remain at CERN and only a small amount will be shipped to RAL for test purposes, debugging, etc.
- All of the ESD (containing reconstruction information), AOD and ETD samples will be required at RAL. Depending on the physics topics studied in the UK, it is conceivable that further filtering of these samples could take place before they are transferred from CERN, but it is premature to assume this and could severely limit the UK physics exploitation of CMS.
- The corresponding simulated event samples will at least double the above numbers. In fact, a raw MC event is estimated to have a size of 2 MB and large-scale Monte Carlo production could easily become the dominant demand on resources (especially if multiple generators, reconstruction requirements, etc, are factored in).
- In the early years at least, several reconstruction passes of the data may be necessary. It is therefore possible that two passes of the data may have to be stored at any one time.

The above leads to the conclusion that any UK system should have the capability of storing of the order of ~500TB/year from CMS. The only practical way to store these data volumes (particularly when the requirements of the other LHC experiments are included) is probably through the deployment of automated tape systems with multi-PetaByte capability and software to stage the essential analysis data onto disk. However, with falling disk prices, it is difficult to reliably estimate the split between disk and tape storage at this stage. A recent IBM announcement of a factor of three improvement in disk density, illustrates the advances being made. For the purpose of modelling, we have, however, assumed the following model with 20% of the data on disk:

	2003	**2005**	**2006**	**2007**
Disk Storage	10TB	100TB	200TB	300TB
Tape Storage	40TB	400TB	800TB	1.2PB
Total Storage	50TB	500TB	1.0PB	1.5PB

3.2 CPU Power

The CPU requirements are less easy to calculate as they depend on the mixture of work and organisational model for performing physics analyses, both of which are largely unknown at the present time. CPU power is measured in SpecInt95 (SI95) units. The conversion factors to other measures are: 1 SI95 = 10 CERN Units = 40 MIPs. The possible computing tasks and their parameters are listed in the following table:

Task	**CPU/event (SI95.sec)**	**No. Events**	**Input Data Volume**
Reconstruction	350 (500 CTP)	10^9	1 PB (Raw)
Selection (Group)	0.25	10^9	100 TB (ESD)
Analysis 1 (Group)	2.5	10^7	1 TB (ESD)
Analysis 2 (User)	3.0	10^7	100 GB (AOD)

One possible model is to assume from 2006 a total of, say, 5 different analysis groups at the 4 CMS(UK) institutes with the following mix of work:

- Event selection performed over 3 days by each group, requiring a capacity of:

  5 groups x 10^9 events x 0.25 SI95-sec/259200 sec = 4823 SI95
- Each group reading the reduced sample of 10^7 events in 1 day, requiring:

  5 groups x 10^7 events x 2.5 SI95-sec/86400 sec = 1447 SI95
- Each physicist submitting individual “user” analysis jobs (4 hour turn-round), requiring a system capable of delivering:

  nphysicists x (1 job x 10^7 events x 3.0 SI95-sec/14400 sec) = nphysicists x 2083 SI95
- Monte-Carlo simulations. The CMS model is for simulation work to be distributed over facilities within the collaboration. Current timings indicate that the generation and simulation of a single event takes 5000 SI95-sec and the digitisation and reconstruction takes 5,800 SI95-sec (under study by Harvey et al). This means that substantial resources will need to be set-aside for this activity.
- It is worth noting that accessing events through Objectivity can introduce a CPU overhead which has been estimated to be ~13%. We have ignored any machines needed for serving data over the storage network between the database and application programs, which may introduce a larger overhead.
- These figures would suggest that a system capable of delivering at least 50,000 SI95 is required for the above mix of work and our model assumes a rough profile of:

	2003	2005	2007
CPU (SI95)	10,000	30,000	60,000
I/O overhead (13%)	1,300	3,900	7,800
Total CPU (SI95)	**11,300**	**33,900**	**67,800**

The figure for 2003 is justified on the basis of immediate simulation requirements (see also section 3.5) as well as the build-up for LHC operation.

For comparison, estimates for CERN and a Tier 1 (20%) Regional Centre, for the period of LHC start-up, have been summarised by Les Robertson as follows:

Requirements for a single LHC experiment	**CERN**		**20% Regional Centre**	
	2005	**2007**	**2005**	**2007**
Processing capacity (SI95)	460,000	910,000	90,000	180,000

The MONARC Architecture Group predicts that the performance of a 4-CPU box will evolve as follows:

	2003	2005	2007
SI95	280	600	1500

From our requirements, this means that the profile for the number of boxes dedicated to CMS(UK) would be something like:

	2003	2005	2007
No. New Boxes	41	38	23

3.3 Effort

It is difficult to be precise about manpower demands because it depends on the precise computing model adopted and the strategies of the other LHC experiments. However, it is possible to define the following potential roles/responsibilities which would be required by CMS:

- Overall management - Liaison with CERN ,etc
- Systems Operation
- Systems Support
- Data (CMS) Management
- DataStore/MSS/HPSS support
- Physics Code Management
- Monte-Carlo Production Management (if undertaken for the collaboration)

For a serious project designed to operate for 15+ years, it is difficult to see how this could be less than 3-4 FTEs in the early years of operation. It is worth noting that a full regional centre is considered to require a minimum of 8 people, provided there is existing infrastructure, which can already be exploited.

On top of this, is the requirement for physicist utilisation of the facility in designing/writing software, analysing physics/detector data, performing simulation studies, etc. This effort is, of course, essential to justify the investment in building a UK computing centre and is the raison d'être for the project in the first place.

3.4 Existing Central Facilities

To put the above estimates into context, the existing disk and tape capacity available centrally at RAL is tabulated below:

Current RAL Particle Physics Data Storage Capacity (circa. Summer 1999)	
DISK SPACE FOR DATA STORAGE	1.2 TB
ROBOTIC TAPE SPACE	12 TB (uncompressed)

Similarly, the total batch CPU capacity at RAL is summarised below: Although there are 100 boxes, these only yield a power of about 1200 SI95 which is shared by all UK experiments.

Current RAL Particle Physics Batch Capacity (circa. Summer 1999)				
Farm	**Machine**	**No. CPUs**	**SI95/CPU**	**SI95**
CSF	HP 712/80	19	2.2	41.8
CSF	HP 735/100	6	2.8	16.8
CSF	HP C110 (PA7200)	4	4.4	17.6
CSF	HP C200 (PA8200)	4	14.3	57.2
CSF	Pentium II (450MHz) – Linux	40	18.0	720
CSF SUBTOTAL		73	-	853.4
NT	Pentium Pro (200 MHz)	10	8.09	80.9
NT	Pentium II (450MHz)	18	18	324
NT SUBTOTAL		28	-	404.9
TOTAL BATCH CAPACITY		**101**	-	**1258.3**

3.5 Shorter Term Requirements

The above sections cover the immediate build-up to LHC start-up and the exploitation of the initial data. However, the requirements for the immediate future, covering the period 2000-2002, also have to be considered. This period is dominated by the need to design and simulate detector components, as well as the production of OO code for the CMS software suite.

In the case of LHC-B, a Monte Carlo Array Processor (MAP) is currently being commissioned at Liverpool after receiving funding through the JREI mechanism of ~£575k. This has the capacity to produce 10^6 to 10^7 fully simulated events per week for the detailed design and optimisation of detector components and physics studies. A complementary case has been prepared to seek funding for an analysis and storage system consisting of forty 1TB disk servers.

This can be used as a basis for CMS(UK) requirements as follows:

- The ability to produce and reconstruct simulated datasets of 10^6 events within one week. The CPU power/event is currently estimated to be ~10,000 SI95-sec. (ATLAS quote a number of 7,000 SI95-sec, which is roughly compatible).
- Storage of such datasets for access by all CMS(UK) collaborators.

The production and reconstruction of one dataset over one week (7 days, 24 hours) would therefore require a system capable of delivering:

10^6 events x 10,000 SI95-sec/604,800 sec = 16,534 SI95

This is roughly consistent with the 11,300 SI95 envisaged above for 2003.

As seen from section 3.4, this far exceeds the current total batch capacity of the available RAL computer farms. A dedicated CMS system approaching this power would therefore need to be installed to fulfil this simulation requirement.

The storage requirement depends on the format of the stored events. The event size for Monte-Carlo data is estimated at 2MB/event, compared to the 1MB/event for real data. If we assume the same expansion factor for ESD format, we arrive at 200kB/event. Each dataset would therefore require:

10^6 events x 2MB = 2TB raw data

10^6 events x 2kB = 2 GB ESD

If we assume that the raw data is required for further reconstruction studies and several datasets will be stored at any one time, then a minimum storage capability of ~10TB would be required, preferably on disk. Again, this exceeds the current capacity of the RAL systems and dedicated space would need to be installed for CMS.

4. Cost Estimation

4.1 Disk Storage

The recent purchase of a RAID array at RAL yields a cost of ~£40/GB for a system based on 36GB disks. Pricing of SCSI disks may be cheaper, but the packaging and extra server requirements have to be included. The Liverpool COMPASS facility is anticipating 1999 costs of only £20/GB for bulk disk purchases. For costing, we average these figures and assume a base figure of £30/GB with a reduction in price every 2 years giving:

	2003	**2005**	**2006**	**2007**
Incremental Disk	10TB	90TB	100TB	100TB
Cost/GB	£7.50	£3.75	£2.82	£1.88
Disk Cost	**£75,000**	**£337,500**	**£282,000**	**£188,000**

4.2 Tape Storage

The costing for robotic tape storage is much more difficult because of the large initial cost needed to purchase robotics and drives – this will be very dependent on the chosen technology, timing of purchases, choice of vendor, etc. It is also a cost that would have to be shared between other experiments.

The prices for the elements of a possible upgrade are given in Appendix 1. These indicate a one-off cost of ~£550K for the purchase of tape robotics, drives and software, which we take to be shared among the three experiments. We assume a 20% share for CMS. No attempt has been made to scale this investment cost from the current (1999) price, on the assumption that a more advanced system would be implemented (this scheme would already saturate capacity within three years of LHC operation, assuming certain technology projections). The recurrent costs amount to ~£130K, again assumed to be shared among the experiments.

	2003	**2004**	**2005**	**2006**	**2007**
Robotics	£550,000	-	-	-	-
Recurrent	£130,000	130,000	£130,000	£130,000	£130,000
CMS Share	**£136,000**	**£26,000**	**£26,000**	**£26,000**	**£26,000**

Media costs are currently ~£600/TB and we assume these to fall by a factor of every 2 years giving CMS a cost profile of:

	2003	2005	2006	2007
Incremental Tape	40TB	360TB	400TB	400TB
Media cost/TB	£150	£75	£56	£38
CMS Media Cost	**£6,000**	**£27,000**	**£22,400**	**£15,200**

4.3 CPU Power

The current Linux upgrade to the RAL-CSF system (using 450MHz, dual-CPU, Pentium II machines) is estimated at £50/SI95. If we assume a factor of 2 reduction in price every 2 years, we arrive at a CPU cost profile of:

	2003	2005	2007
Incremental CPU (SI95)	11,300	22,600	33,900
Cost/SI95	£12	£6	£3
CPU Cost	£135,600	£135,600	£101,700

The PASTA99 Working Group has recently predicted CPU costs in 2005 of $10/SI95 for PCs and $20/SI95 for Enterprise-like servers. These costs would appear to be in-line with our assumptions.

Pulling all the costs together, yields the following initial profile:

	2003	2004	2005	2006	2007	TOTAL
Disk	£75K	-	£337.5K	£282K	£188K	**£882.5K**
Tape Robotics	£136K	£26K	£26K	£26K	£26K	**£240.0K**
Tape Media	£6K	-	£27K	£22.4K	£15.2K	**£70.6K**
CPU	£135.6K	-	£135.6K	-	£101.7K	**£372.9K**
TOTAL	**£352.6K**	**£26.0K**	**£526.1K**	**£330.4K**	**£330.9K**	**£1566.0K**

4.4 Development costs and JIF

The costs outlined above amount to a total of roughly £1.57M, spread fairly evenly over the years 2003-7. This assumes that expenditure starts in 2003. In our final model, we assume that we start the development earlier, building up to the performance outlined above for 2003.

For disk and CPU purchases, we allow for a constant annual expenditure between 2000-3. Taking into account the assumed improvements in price/performance, this means that we purchase about 14% of the required capacity (for 2003) in the year 2000, increasing to ~40% in the year 2003. The purchase of tape robotics equipment is assumed to be a one-off item in 2003. The total cost up to 2003 becomes £614k, as opposed to ~£350k, and the overall project cost increases to £1.96M. As argued above, this extra expenditure buys us early access to the computing facility for physics and trigger studies, and the development of reconstruction code. The equipment purchased in 2000 is equivalent to about twice the current available CPU power and 35% more than available disk capacity.

The expenditure profile now breaks down as follows (in £k):

	2000	2001	2002	2003	2004	2005	2006	2007	TOTAL
Disk	33	33	33	33	0	338	282	188	**940**
Tape Robotics	0	0	0	136	26	26	26	26	**240**
Tape Media	0	0	0	6	0	27	22	15	**70**
CPU	85	85	85	85	0	136	136	102	**714**
TOTAL	**118**	**118**	**118**	**260**	**26**	**527**	**466**	**331**	**1964**

5. Conclusion

The computing requirements and effort to base CMS physics analysis in the UK go far beyond the resources likely to be available within the individual institutes. The only tenable approach is to base large-scale UK data analysis at a central facility. We have looked at a possible set of specifications for such a facility. Based on these ideas, we have produced a first rough estimate of £1.65M capital costs over the years 2003-7 for CMS.

However, it is felt that the three experiments, ATLAS, CMS and LHC-B, would benefit from a common LHC facility to avoid duplication of effort and to maximise the efficient use of resources. We argue that there are benefits to be gained from beginning work earlier using JIF funding, and it is suggested that we submit a joint JIF bid to fund such a facility. Our model suggests that the CMS requirements are around £614k plus staff costs, with about twice this sum to be found from other sources after the end of JIF.

The support of all the CMS(UK) institutes is now needed to progress towards a well founded project. In addition, we encourage feedback on the various parameters and models, which have been assumed in this document, with a view to firming up the CMS requirements for wider discussion.

APPENDIX 1 : ROUGH COSTINGS FOR ELEMENTS OF AN AUTOMATED TAPE STORAGE SYSTEM

For mass storage, RAL currently has an IBM 3494 Tape Library Dataserver with five 3590 (Magstar) drives and 2,900 slots. Using the current 3590 (10GB) cartridges, this gives a total capacity of 29TB, of which particle physics is allocated 12TB.

In addition, there is a "mothballed" 5,500 slot StorageTek ACS robot which is currently only equipped with 3480 drives and tapes. An identified route to meet initial LHC storage demands is to upgrade this silo with Redwood drives and D3 tapes. Using the current 50GB tapes, this would give a total capacity of 275TB, but the projections are for 1TB+ tapes by 2005(?), giving a total capacity of around 5.5PB.

An alternative, is to use the recently released 9840 ("Eagle") media by STK. The drives are cheaper (£25K), but the tapes are more expensive and provide a faster retrieval time. Unfortunately, the development profile predicts a tape capacity of only 140GB by 2003. A mix of Redwood and 9840 media may provide a possible solution, but for the purpose of costing we assume complete population by Redwood drives and media.

Capital Costs

5*Redwood drives (£60K each)	£300K
5,500 tapes (£30 each - current 50GB flavour)	£165K
STK Control software (ACSLS)	£30K

Recurrent Costs

Maintenance costs (current)	£27K/year
ACSLS support	£6K/year

Some of these costs would be shared with other experiments.

It is possible that new tape technologies will emerge before 2005. IBM recently announced a laboratory demonstration of a 100GB cartridge using the Linear Tape Open (LTO) format and products are planned for late 1999.

Hierarchical Storage Manager

A number of studies and projects are underway to identify a hierarchical storage manager for particle physics in the LHC era. The main commercial system is HPSS, which has been developed by IBM and a consortium of mainly US laboratories. HPSS is deployed at FNAL, SLAC and IN2P3. CERN has been evaluating and developing an HPSS system for upcoming experiments like CHORUS and NA57, but have put off a final decision on what to choose for LHC experiments for another 1-2 years.

Alternatives include the Eurostore project, which DESY will employ and the CASTOR project, which is an upgrade to the SHIFT staging software.

For the purpose of costing, we currently assume HPSS, which is expected to be the most expensive system.

HPSS licence costs	$300K first year $150K subsequent years (if acquired in US, otherwise 20-25% premium in Europe)
Other costs (eg. DCE, Encina, Sammi, etc)	$50K

Again, these costs would be shared with ATLAS & LHC-B, if there were a common strategy.