Midwest ATLAS Tier 2 Renewal Proposal for 2012-2016

University of Chicago
Indiana University

1. History and Status of MWT2

The MWT2 project began in 2005 with the first production deployments in the summer of 2006. At each step the MWT2 has evolved according to principles established by the ATLAS computing model as well as practical lessons from the early data challenges and functional tests, production campaigns, and from experience with analysis workloads during the first year of the LHC physics program. Table 1 shows the deployed resource profile over the past five years.

	2005-06	2006-07	2007-08	2008-09	2009-10
MWT2 Original Proposal					
CPU (HS06) "Total Dedicated"	2046	3581	4516	5071	NA
Storage (TB) "Total Dedicated"	157	326	530	855	NA
ATLAS MOU Values for MWT2					
CPU (HS06)		3304	4448	4960	11040
Storage (TB)		213	282	520	1060
Delivered Resources at MWT2					
Job-slots	312	623	1604	2352	3628
CPU(HS06)	2230	4068	12587	17508	30688
Useable Storage (TB)	68	226	524	1470	1470
Added HS06	2230	1838	8519	4921	13180
Added useable TB	68	158	298	946	0

Table 1: The blue shaded figures include capacities from the spend-out of the FY10 funds we expect to be in operation in the current fiscal quarter. During FY11 we will retire some of the equipment from the 2006 procurement.

The delivered resources include substantial equipment and infrastructure contributions from our respective universities as we have outlined in the associated cost-benefit narrative for MWT2. This has allowed MWT2 to become a major production and analysis resource for the ATLAS physics community, consistently ranking in the top 5 of 163 WLCG Tier2 federations in terms of delivered HS06-hours and storage capacity.

In addition to this production infrastructure, the MWT2 team has contributed test-bed resources, troubleshooting support, and operational feedback throughout the long period of OSG software stack development, ATLAS Panda and DQ2 development, and storage system deployments (primarily dCache but also some xrootd for Tier3 sites).

2. MWT2 Infrastructure

2.a Existing Facilities

The MWT2 is hosted at two sites: the Enrico Fermi Institute at the University of Chicago and at the Indiana University Informatics and Communications Technology Complex (ICTC) in downtown Indianapolis. Both universities have provided the space, power and cooling for the past five years as well as the network infrastructure and regional provider connectivity fees. Our cost-benefit narrative provides details on these expenses, which have been substantial. These expenses will continue to be provided for next five-year period.

At the University of Chicago, the machine room is an 850 square-foot space located next to the high energy physics building with an 18” raised floor, two Liebert CRAC units (50T, 175 kW), and a 40 kW UPS system. All storage and head-servers are backed up by the UPS system. Compute nodes are powered from directly from the wall service. During the second year of the project a 10 Gbps network service was provided to the room. At present the MWT2 occupies 15

racks with space, power and cooling available to cover FY10 spend-out. Future capacity will be provided by a major university investment in a new server room, scheduled for spring/summer 2011 as described below.

At Indiana University, the ICTC is a 213,815-square-foot, $43.6 million building on the Indianapolis campus. The building houses a hardened 8,400 square foot machine room, equipped with uninterruptible power supplies for all computing systems, backed up by a 1.25 MW generator. Several major IU computing projects, including the IU Teragrid resource and the OSG GOC are co-located here, and it is a major peering hub for Internet2. At present MWT2 occupies 6 racks in this space with room to expand to 10 as resources for hardware purchases become available.

2.b Facility Expansion

At Chicago the university is building a new physical sciences research center that will replace the structure currently hosting MWT2 equipment. Cognizant of our long-term commitments to the ATLAS computing program, the Dean of the Physical Sciences Division built into the project plan construction of a new server room for the Fermi Institute, with ATLAS as the primary stakeholder at a cost of $1.2M from an alumni gift. The new server room will be located in the basement of the adjacent "accelerator building" that once housed Fermi's proton cyclotron (and as such has an ample 2400V power feed). The 2000 square-foot room will have a 750 kW electrical service for both computing and mechanical, with 360 kW for computing equipment of which 140 kW will be backed by UPS. Space for up to 40 racks for ATLAS will be available though we wouldn't see the need to expand to that number of racks in the next five years.

At Indiana there is also ample space, power and cooling for expansion. While the existing reserved space for ATLAS is 10 racks, it is highly like additional space and associated infrastructure would be available should we request it.

2.c Network

The network path between the two sites is 10 Gbps end-to-end path over two regional providers: iLight connects Indianapolis to Starlight in Chicago while iWire connects UC to Starlight, peering in the MREN (Metropolitan Research and Education Network) switch at 710 North Lakeshore Drive. A VLAN is configured along the path so as to directly route traffic between the two subnets. Connectivity to Internet2, important for Tier 3 access, is provided by the CIC OmniPoP router at Starlight. CIC (Committee on Institutional Cooperation) membership includes a number of institutions that peer at various points in downtown Chicago, including UC and most of the Big Ten universities, in particular: IU, UM, MSU, Illinois and Iowa. UC has a second 10 Gbps lambda that runs from the campus edge in Hyde Park up the Illinois Central tracks to downtown where it peers directly with ESnet, providing at direct path to Brookhaven over a dedicated circuit maintained by ESnet. The UC campus and edge networks are being redesigned with Nexus 7000-series switches to provide a 40 Gbps core in the next 2 years, with possible expansion to 100 Gbps in the next 5-years as needed. Details are provided in the appendix.

3. Personnel

MWT2 has a shared administrative model: the privileged systems administration domain of either facility is available to all systems staff so as to provide redundancy in operational response and systems monitoring, and to share expertise between the sites thereby achieving some measure of administrative efficiency.

The primary systems administrator at UC is Aaron van Meerten who is funded 100% by the Tier2 project. He is backed up by Nate Yehle who is funded at 33% by the Tier2 grant. At the present time Nate is funded partly by OSG and is also responsible for administrating the ITB site at UC. When OSG funding ends, the remaining 67% salary will be provided by the university.

At Indiana the primary administrator is Sarah Williams who is funded by the Tier2 project at 67% with the university paying 33%. A half-time technician, Rakesh Kaupu, is paid by the Tier2 grant. Fred Luehring has been supported at .15 FTE for his role in management oversight for the IU site. Fred additionally has been an important contributor in several US ATLAS facility-wide tasks, most recently validation of all the Tier2's for proper Squid-Frontier caching of ATLAS conditions data. For the 2012 to 2016 grant period his percentage will be reduced to 5% from the Tier2 grant.

Charles Waldman was funded as a Tier2 administrator for his first two years of employment at UC. He has since been transitioned to work on common US ATLAS Facility projects dealing with Tier2 data management, including consistency checking between the DQ2 central catalog, the LFC catalog and Tier 2 storage systems. He has delivered several tools to the community including scripts to clean up the proddisk area, the local site mover tools, and worker node caching functions. He has collaborated extensively with several members of the US ATLAS collaboration and has interacted often with the Panda development team, especially as regards to pilot issues and Python programming. More recently he has worked on data access issues focusing on direct dcap access, both local and wide-area, providing input back to the dCache development team. His current focus is on providing an xrootd federation infrastructure, providing a new layer of access to Tier2 facility storage systems.

4. Division of Resources

For the first five years of the project we have divided the resources evenly between the two sites at $300K/year per site, with one exception in FY09 when we placed more storage at UC so as to couple to the MWT2 analysis queue. In the next five year period we will divide resources along the lines of 70%/30% split between UC and IU but we will remain flexible in provisioning hardware and will move funds as needed to optimize local resource contributions.

5. Resource Commitments

We have secured the following commitments from our respective university administrations:

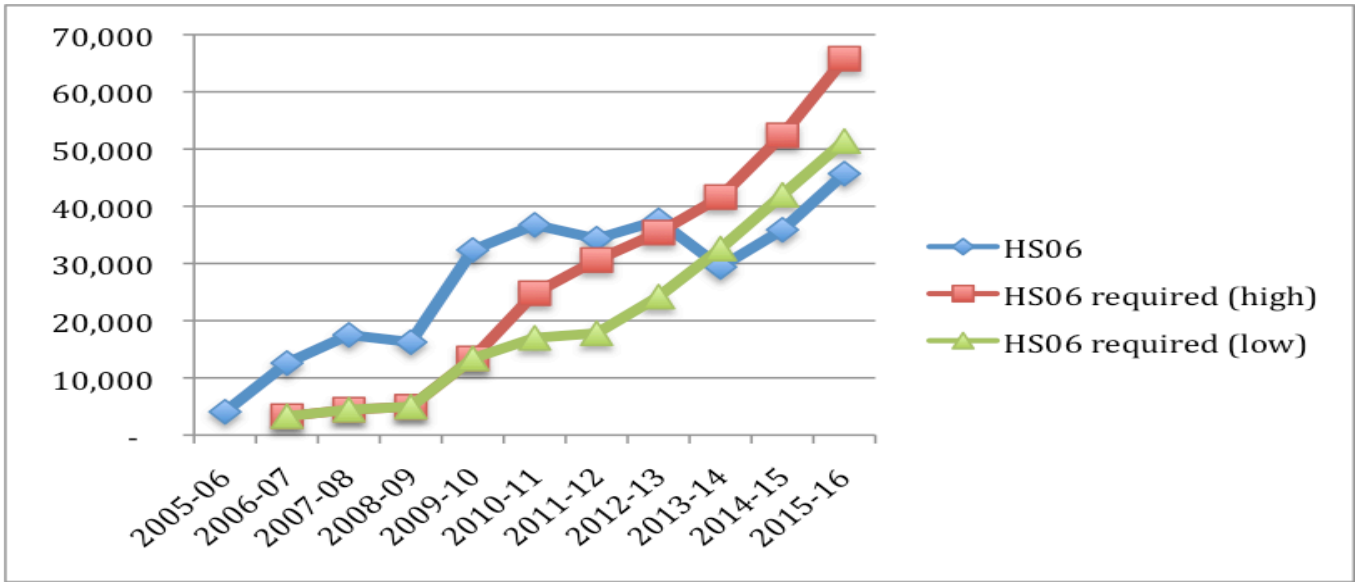
1. Construction of a new server room at UC with 40 racks reserved for ATLAS
2. Salary and benefits for 2/3 FTE equivalent at UC
3. $35K/year towards salary and benefits at IU, equivalent to > 33% FTE

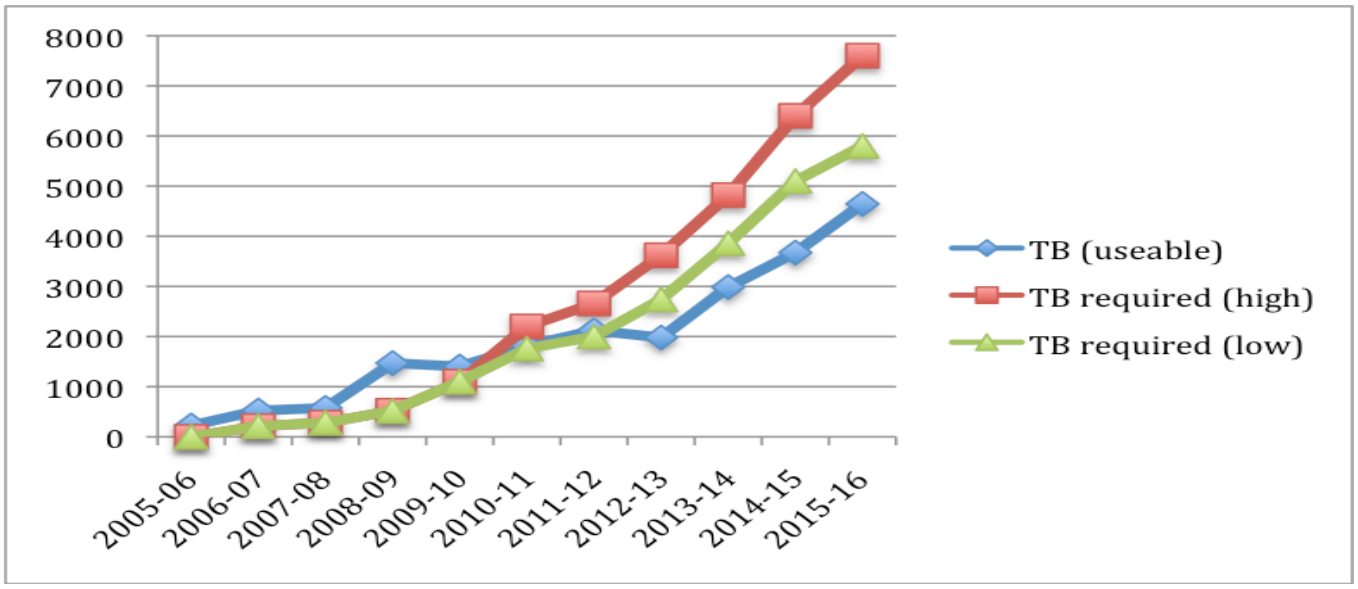
6. Spending Plan ($600K/y with 1.5% growth)

The table below assumes the parameters discussed during the cost-benefit phone meetings. The "Network, headnodes, etc" category covers additional equipment needed to provide the computing and storage, for example network switch modules, cables, racks, PDU, etc. In estimating this cost we have made some simplistic but likely conservative assumptions: adding two 42U racks per year on average (there will be a year-to-year variation due to retirements), $27/GigE port for switching with servers dual-cabled, and costs for both 10G and GigE network cables. We list explicitly dollar university contributions that are applied to effort making up the 2/3, 1/3 FTE contributions, and have left a placeholder for equipment funds from UC we may be able to add during the project.

Projections	AVG burdened effort:		$363,752		
$600K funding scenario	2011-12	2012-13	2013-14	2014-15	2015-16
Personnel (project paid)	$342,572	$352,849	$363,435	$374,338	$385,568
Personnel (UC paid)	$56,546	$58,242	$59,990	$61,789	$63,643
Personnel (IU paid)	$35,000	$35,000	$35,000	$35,000	$35,000
Network, headnodes, etc	$25,651	$25,651	$25,651	$25,651	$25,651
Total "non-equip" (project)	$368,223	$378,500	$389,085	$399,988	$411,218
Equipment (600K budget)	$217,777	$230,500	$229,050	$227,419	$225,600
Equipment supplement (UC)	$-	$-	$-	$-	$-
compute %	40%	40%	40%	40%	40%
storage %	60%	60%	60%	60%	60%
CPU $	$87,111	$92,200	$91,620	$90,968	$90,240
Disk $	$130,666	$138,300	$137,430	$136,451	$135,360
CPU $/HS06	$14.29	$11.34	$9.00	$7.14	$5.67
Disk $/TB	$216.68	$171.98	$136.50	$108.34	$85.99
CPU HS06 purchased	6,097	8,131	10,180	12,735	15,916
CPU HS06 removed	(8,519)	(4,921)	(18,333)	(6,233)	(6,097)
Disk TB purchased	603.0	804.2	1,006.8	1,259.5	1,574.1
Disk TB removed	(298.0)	(946.0)	-	(572.2)	(603.0)
Job slots purchased	677	903	1,131	1,415	1,768
Job slots removed	(1,293)	(520)	(2,037)	(693)	(677)
HS06	**34,324**	**37,535**	**29,381**	**35,883**	**45,702**
TB (useable)	**2,121**	**1,979**	**2,986**	**3,673**	**4,645**
Job slots	3,615	3,999	3,093	3,815	4,906
HS06 required (high)	30,600	35,400	41,600	52,400	65,800
HS06 required (low)	17,800	24,200	32,600	42,000	51,400
TB required (high)	2,660	3,620	4,820	6,400	7,600
TB required (low)	2,000	2,740	3,860	5,100	5,800
University contributions	$91,546	$93,242	$94,990	$96,789	$98,643

With this spending plan we have the below comparison with required high/low pledges. The installed capacities are the blue curves. Clearly the CPU provisioning is in good shape until 2013-2014 where it falls below the low requirement. On the other hand with these assumptions we would expect shortfalls in storage beginning in 2012-2013.





7. Appendix

7.a Network Upgrade Plan

The University of Chicago's current network consists of the three-tiered model: core, distribution, and access layers. The capacity between the distribution and core layers is primarily 1 Gbps redundant links with a few 10 Gbps links to our research facilities on campus such as the MWT2 machine room. The uplinks from the access to the distribution layers primarily consist of multiple 1 Gbps links. Redundant Internet2 connectivity terminates into the core switches via two 10 Gbps Ethernet circuits.

Over 3 years, the entire network will be redesigned. This will encompass the above tiers as well as add a network Edge. The Distribution layer will entail a hardware component upgrade on the existing switch. This will enable us to implement Virtual Switching System (VSS) that provides Multi Chassis Ethernet Channel (MCEC) and in-service software upgrade capability. The access layer is transitioning to a standard Cisco loop-free model, by using stackable switches and using

MCEC to the Distribution Layer. The current core switching will be replaced by the Nexus7000 hardware. The Nexus hardware upgrade future proof's the network by laying the foundation for the initial 40 Gb then later 100 Gb core network.

A new edge is being installed in order to protect the core network from external threats. It will also offload internet routing tables from the core switches. The new edge routers are Cisco ASR1000's. This design enables implementation of Cisco's Performance Routing (PfR) in order to optimize traffic flows by application. Based on this it allows us to shift traffic to circuits with lower utilization in order to protect applications deemed as priority.

The bandwidth between all distribution and core locations will be upgraded from the current mix of 1 and 10 Gbps circuits to four 10 Gbps circuits using MCEC and thus providing a combined 40 Gbps throughput to each distribution location. 40 Gbps processor cards enable us to install 40 Gbps Ethernet when made available by our Regional Optical Network (RON). Once our Internet2 RONs have greater than 10 Gbps port capacity, we will be able to connect at those speeds. The Cisco ASR product roadmap enables an easy upgrade to 100 Gbps within the next 3-5 years.

The MWT2 server room will be connected to a distribution site via two 10 Gbps MCEC circuits, providing a combined 20 Gb throughput initially. Today it's limited to a single 10 Gbps uplink because its redundant link is blocked via Spanning Tree. In the future additional capacity can be added by increasing the amount of 10 Gbps uplinks.

UChicago's ITServices partnered with Stanford University, Indiana University and Google and submitted a GENI Racks proposal in response to GENI Solicitation 3 in Aug 2010. GENI-racks are the basic unit of computation and disk storage within the larger GENI facility. The solicitation requires 40 racks to be deployed over 3 years with at 8 of these deployed during the first year. The racks will enable research in cloud computing, data center networking, content distribution and management, in-network processing and so on in conjunction with the OpenFlow based GENI network substrate. ITServices is working with our RON to enable the Internet2 ION virtual circuit network service that provides dedicated bandwidth to researchers.