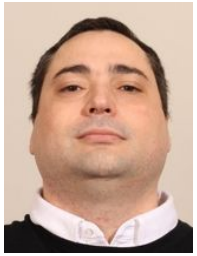
US ATLAS Computing Facilities Face-to-Face
1 December 2022 - SLAC

NET 2.1 Discussion

Rafael Coelho Lopes de Sa, Verena Martinez Outschoorn, **Eduardo Bach**, Will Leight

Introduction to the team

UMass Physics NET2.1 Team



Eduardo Bach
Sys Admin

Will Leight
Sys Admin

Rafael Coelho Lopes de Sa
Faculty & PI

Verena Martinez Outschoorn
Faculty

Harvard Physics & Research Computing

John Huth
Physics Faculty

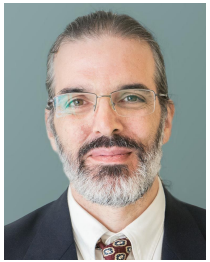
Scott Yockel
RC Head

Milan Kupcevic
NESE Lead Engineer

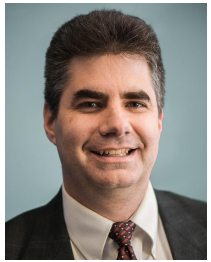
Brian White
Manager of sys. eng. & ops.

UMass Research Computing

John Griffin
Sys Admin
campus cluster



Rick Tuthill
Interim RC Head



Jim Mileski
CTO

Chris Misra
VC & CIO

Others at MGHPCC

John Goodhue
MGHPCC director

Christoph Paus
MIT faculty

Max Goncharov
MIT Sys Admin

NET 2.0 Equipment Inventory & NET 2.1 Purchase Plan

An inventory of equipment purchased by BU was requested. Only partial information was provided.

- BU has provided an inventory of computing nodes. It approximately matches the receipts received by Stony Brook University.
- Harvard has provided an inventory of NESE storage nodes and DTNs. It matches the receipts received by Stony Brook University.
- BU has not provided an inventory of network equipment.

The agreement between US ATLAS, BU & UMass upper management is that BU would continue until early 2023 (~end of January) **to assist in the transfer**.

NET 2.0 Equipment Inventory & NET 2.1 Purchase Plan

Equipment	Number of servers	Slots	Equipment year	Total kHS06	Phase
AMD EPYC 7302 16-Core Processor	40	2560	2021	43	1
AMD EPYC 7302 16-Core Processor	39	2469	2021	42	2
AMD EPYC 7302 16-Core Processor	8	512	2021	7.4	3
Intel(R) Xeon(R) Gold 6148	24	1920	2019	20	3
AMD EPYC 7542 32-Core Processor	16	1024	2020	15	3
Miscellaneous leftovers					4
Total	127	8512		127	

1. Phases 1 and 2 each correspond to a BU rack. It makes the transfer easier.
2. We have space for phase 1 and 2. We do not have more space **at this moment**.
3. We have already requested the first rack more than 2 months ago. Neither BU nor UMass IT has done it. We do not have control over their schedule beyond continuing to ask.
4. **We need everyone in US-ATLAS to tell BU that transferring the equipment is the priority**

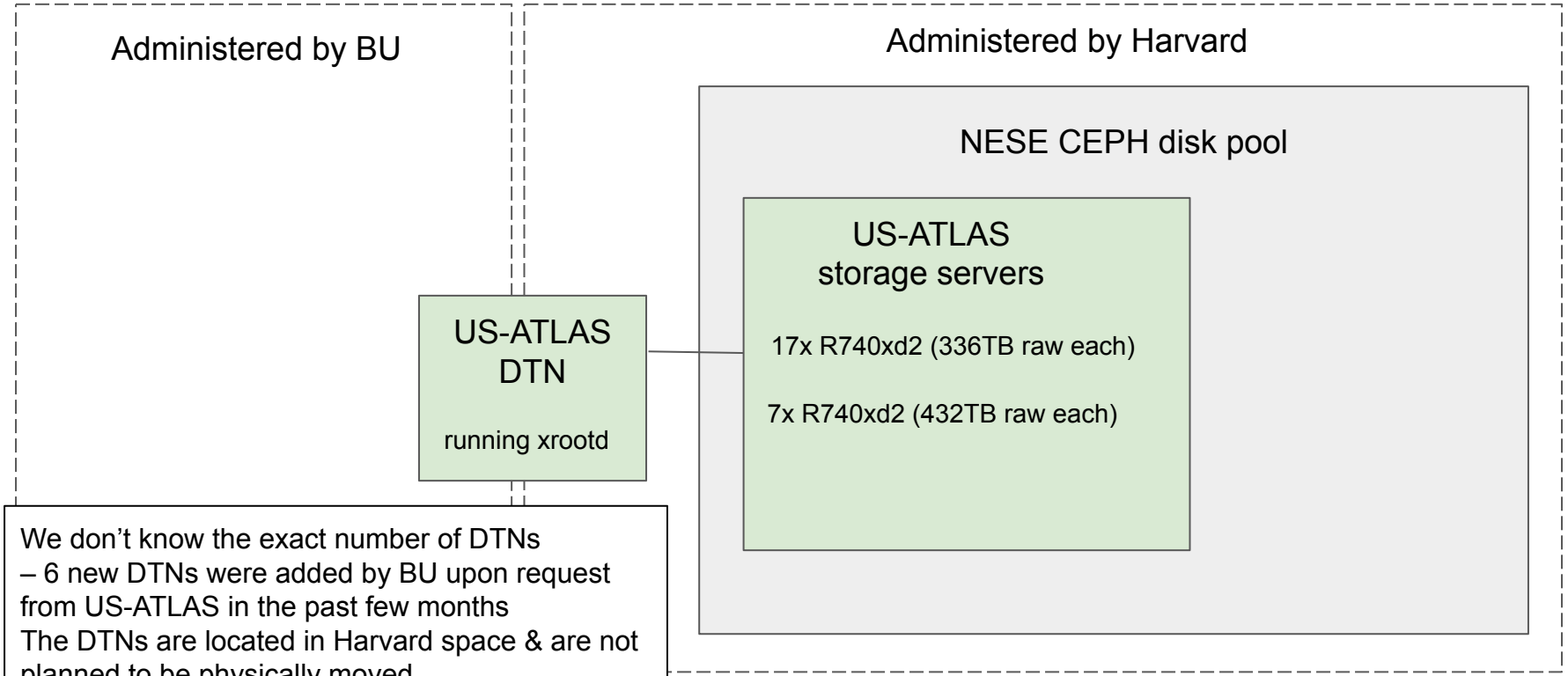
NET 2.0 Equipment Inventory & NET 2.1 Purchase Plan

Resource	Available from BU	Pledge 2022	To be deployed	Pledge 2023	To be deployed
Computing [kHS06]	127	81	162	89	178
Storage [PB]	6.6	7.2	8.7	8.5	10.3

Necessary procurement for 2023:

We have a deficit of 178-127 = 51 kHS06
Estimated cost for 2023 ($5/HS06): $255,000

Current NESE structure



We don't know the exact number of DTNs
– 6 new DTNs were added by BU upon request from US-ATLAS in the past few months
The DTNs are located in Harvard space & are not planned to be physically moved
We need to transfer the ownership & administration from BU to UMass

Current NET2 NESE Equipment

Equipment	Number of servers	Equipment year	Total raw [PB]	Total usable (EC 8+3) [PB]
Dell PowerEdge R740xd2	17	2021	5.7	4.1
Dell PowerEdge R740xd2	7	2022	3.4	2.5
Total	24		9.1	6.6

NESE cost model

Item	Units	Size
Storage server cost	$25,837.00	402 TB (raw)
Storage server housing	$3,004.00	
Current servers	17	4.1 PB
	7	2.5 PB
NET2 to be deployed (2023)		10.3 PB
Additional storage needed		3.7 PB
Additional storage needed (raw)		5 PB
Additional number of servers	13	
Total number of servers	37	
Server cost (2023)	$335k	
Yearly housing cost	$111k	

1. It does not take into account DTN costs
2. Yearly maintenance cost is prohibitive
3. Total estimated procurement for 2023: $255k + $335k = $590k (~**$400k estimated for 2023**)

Proposed NESE structure

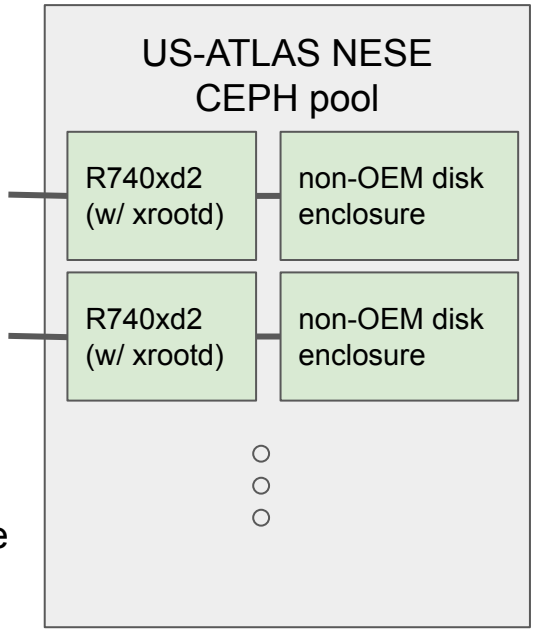
Administered by US-ATLAS (with Harvard)

Instead of paying the yearly cost we would have a separate slice of NESE administered by US-ATLAS (with Harvard).

Separate slice could operate with even less aggressive EC than 8+3.

Add non-OEM disk enclosures to reduce cost per TB.

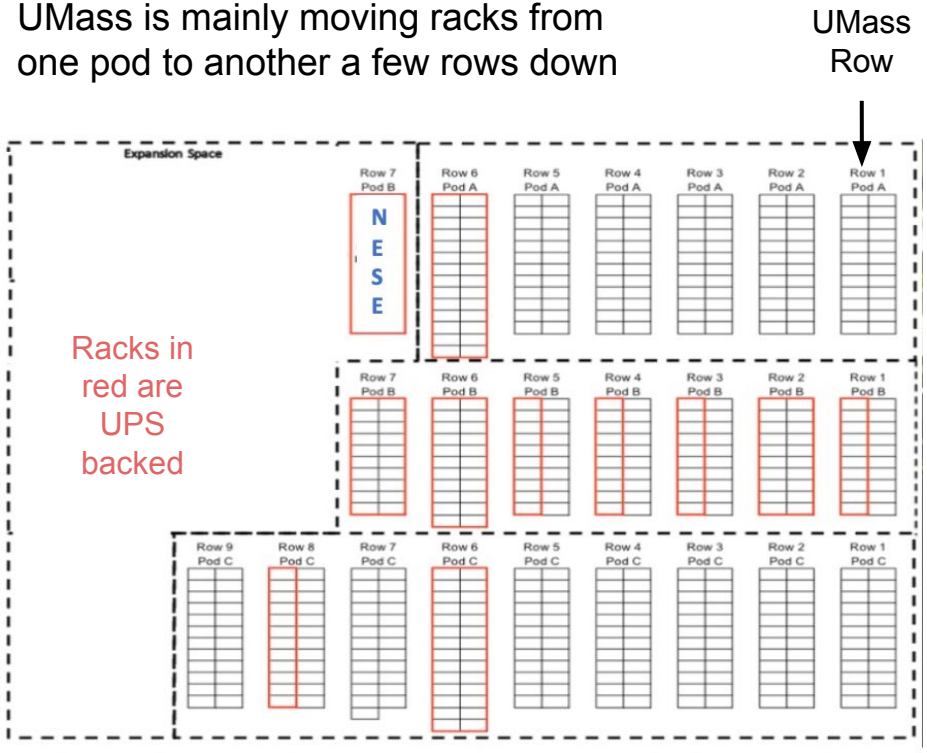
This would make the cost per usable PB closer to other Tier 2s.



NESE CEPH disk pool

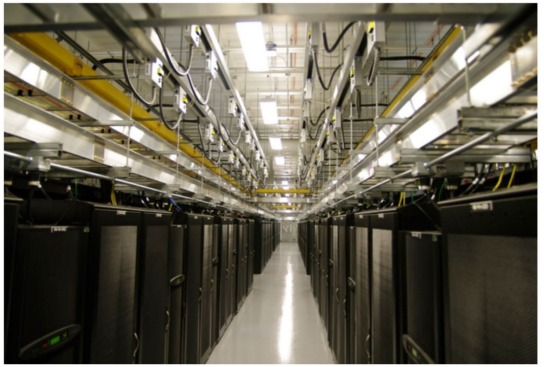
MGHPCC Space Overview

The transfer of equipment from BU to UMass is mainly moving racks from one pod to another a few rows down

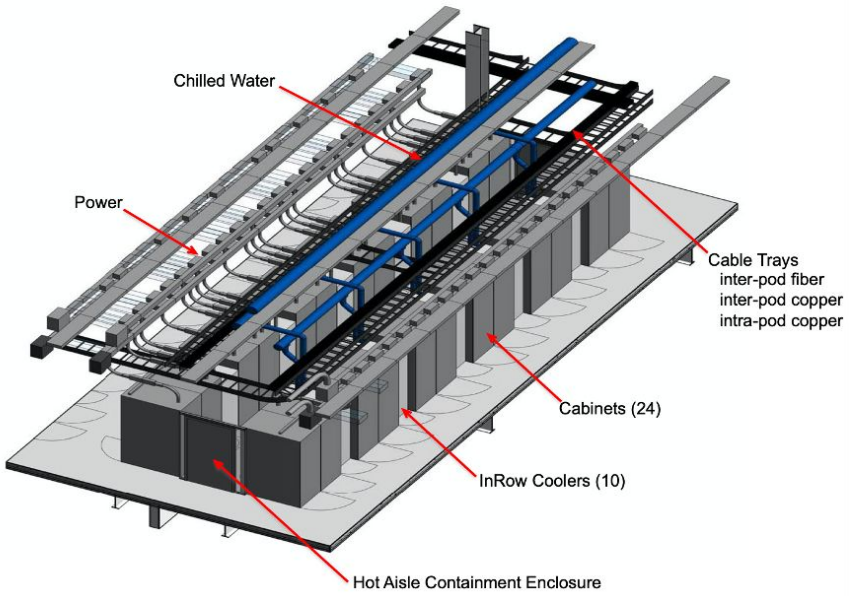


Each of the MGHPCC institutions originally invested in one row with three pods for each institution

- Rows 1-5 for UMass, Harvard, MIT, BU and Northeastern
  – Row 1 is UMass space
- Space for expansion has started to be used, especially by MIT and Harvard
- NESE occupies a pod in row 7



MGHPCC current rack availability



Currently we have 2 racks available in a UMass pod, which should be enough for the first two transfers.

The transfers were first requested more than 2 months ago, but UMass IT and BU IT have not yet moved the equipment.

Current racks use cooling every three racks and can dissipate up to 15 kW, which is not enough to operate all the machines we will receive in each of the phases.

Additional racks may be made available in the current UMass pod, but it is not clear that US-ATLAS would have all of them.

Using 15kW is not ideal, since only a small fraction of the rack would be used.

MGHPCC upgrade options

We are investigating 2 options for upgrade of the MGHPCC

1. Upgrade an existing UMass pod to backplane cooling which would add 8 racks of space for US-ATLAS from space gained from cooling
2. Build a new dedicated pod for the Tier 2 with 8 new racks

Both options use rear door heat exchangers which are more space efficient and do not require hot aisle containment. The new racks support 30kW. Upgrade takes 4-6 months, driven by long lead times for a few parts

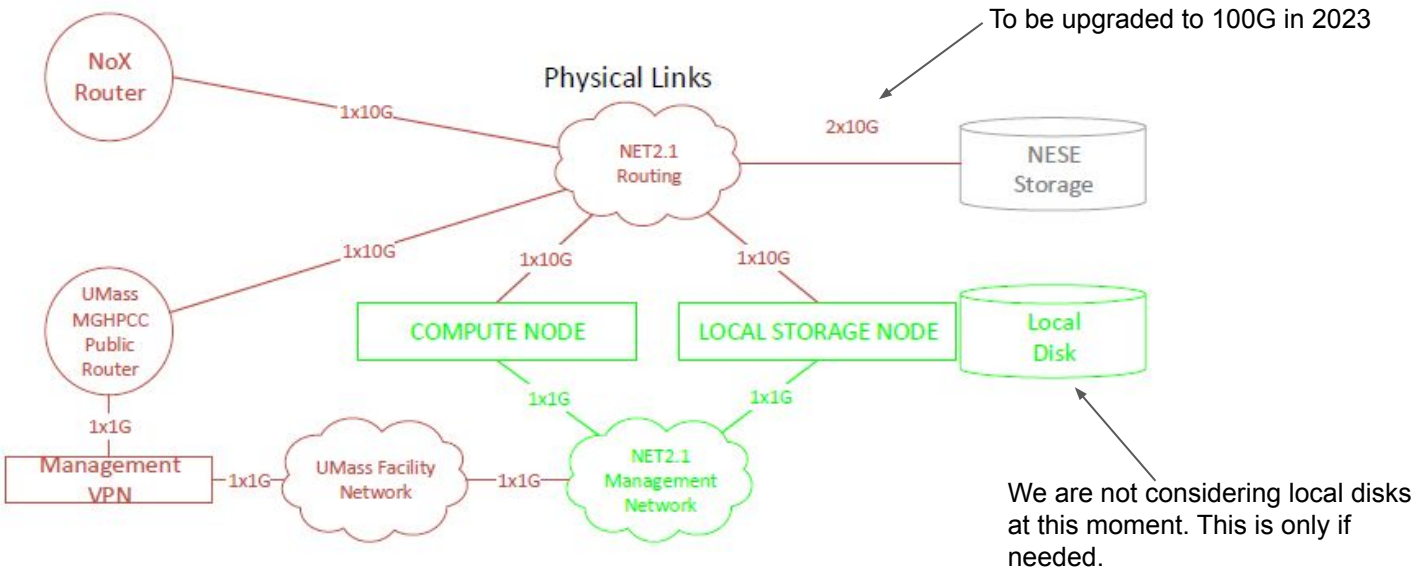
Item	**Cost**
Infrastructure	300k
8 Racks	160k
Total	**460k**

The current BU equipment fits in 4 racks (30kW) – motivates the transfer in 4 phases.

A total of 8 racks would allow sustainable growth for the equipment in the next years.

MGHPCC networking

1. Basic internal (MGHPCC) network infrastructure is currently being put in place for NET2.1
2. All the bandwidths marked are for 2022 and will be updated in 2023 as needed/possible.
3. We requested this connection in October, it is progressing but not ready yet: there is not much incentive to install the network if BU is not transferring equipment.



ESnet connection

Current BU connection

This connection costs $100k per year.
UMass has no other project using an ESnet connection at this moment
No interest from the university to pay such large costs.

Preliminary UMass NET2.1 connection

This connection costs $20k per year. UMass Physics will pay in 2022.
Not sustainable in the long run, but will give us time to look for alternatives.

Possible future ESnet connections

1. We could have a shared 100 Gbps connection with MIT via NoX (MIT owns NoX). But that would imply a collaboration between the two Tier 2 systems (ATLAS – CMS/LHCb), which would have non-negligible implications for the system.
2. ESnet has a presence at 300 Bent St. in Cambridge. UMass has a presence at 1 Summer St. in Boston. Infrastructure investment could connect both but an agreement would need to be reached with UMass.
3. New regional network (NEREN). Agreement and infrastructure investment with NEREN could create a connection between MGHPCC and ESnet at either NYC or Boston.

All of these options are very open at this moment and require negotiations with UMass, NoX, NEREN and other stakeholders in New England.

NESE tape and possible collaboration with MIT

1. NESE tape is fully functional. DTNs have xrootd endpoint installed and the connection with CERN has been tested.
2. Integration with ATLAS-ADC has been paused until transfer of equipment from BU happens, but CMS is going ahead with it.
3. MIT has demonstrated interest in having a common Tier 2 system at MGHPCC so that both could operate and benefit from NESE tape with support from Harvard.
4. A collaboration with MIT of this sort could be a solution for the ESnet connection (since MIT owns NoX).

We are requesting a VLAN between our ATLAS Tier 2 and the CMS/LHCb Tier 2 at MGHPCC to do some preliminary network studies.

We are planning to continue to work on NESE tape in the future. The ATLAS group at Harvard is also interested in contributing together with their research computing team that already manages NESE.

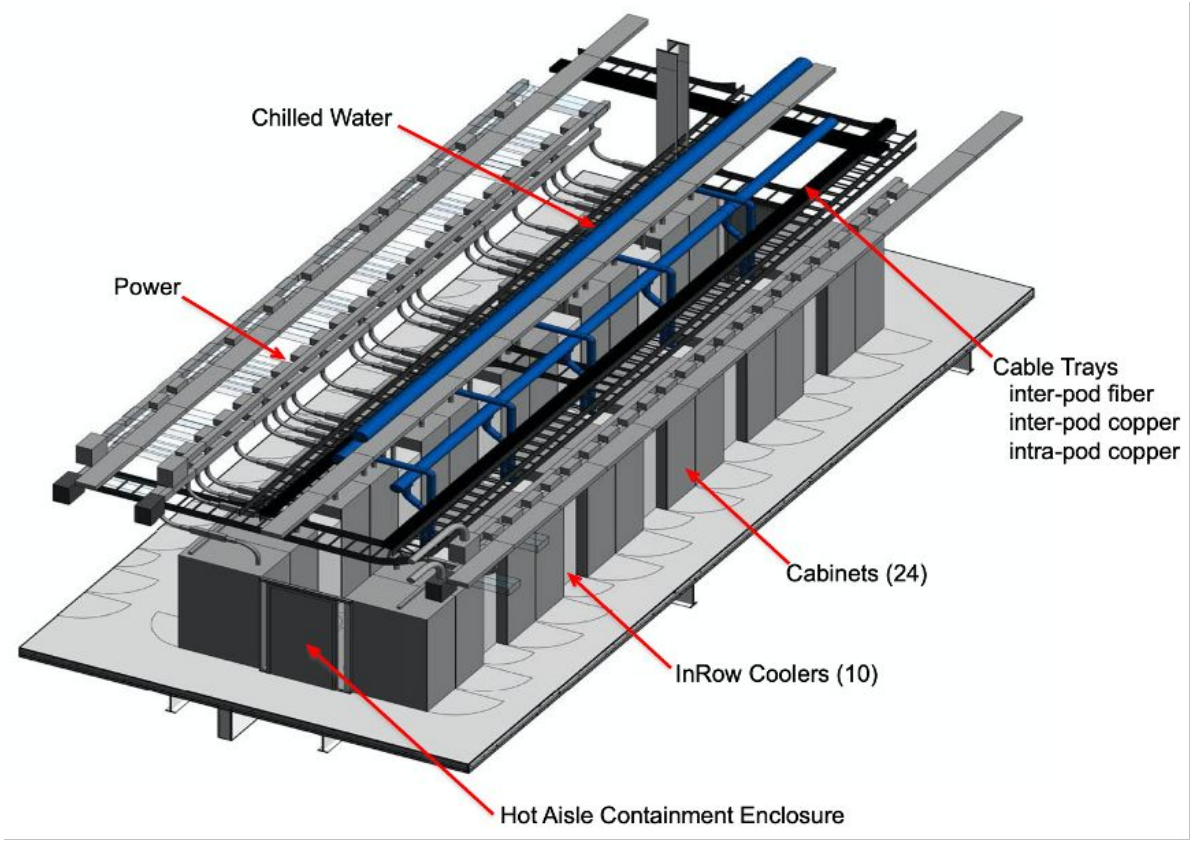
NET2.1 as a native Kubernetes cluster

- Containerization is one of the R&D projects US-ATLAS is interested in undertaking in 2023.
- We are interested in exploring this possibility at NET2.1, since we are starting a new system basically from scratch.
- NET2.1 can provide additional experience and know-how in this area.
- In the long run, kubernetes can also provide a more streamlined infrastructure with easier deployment.

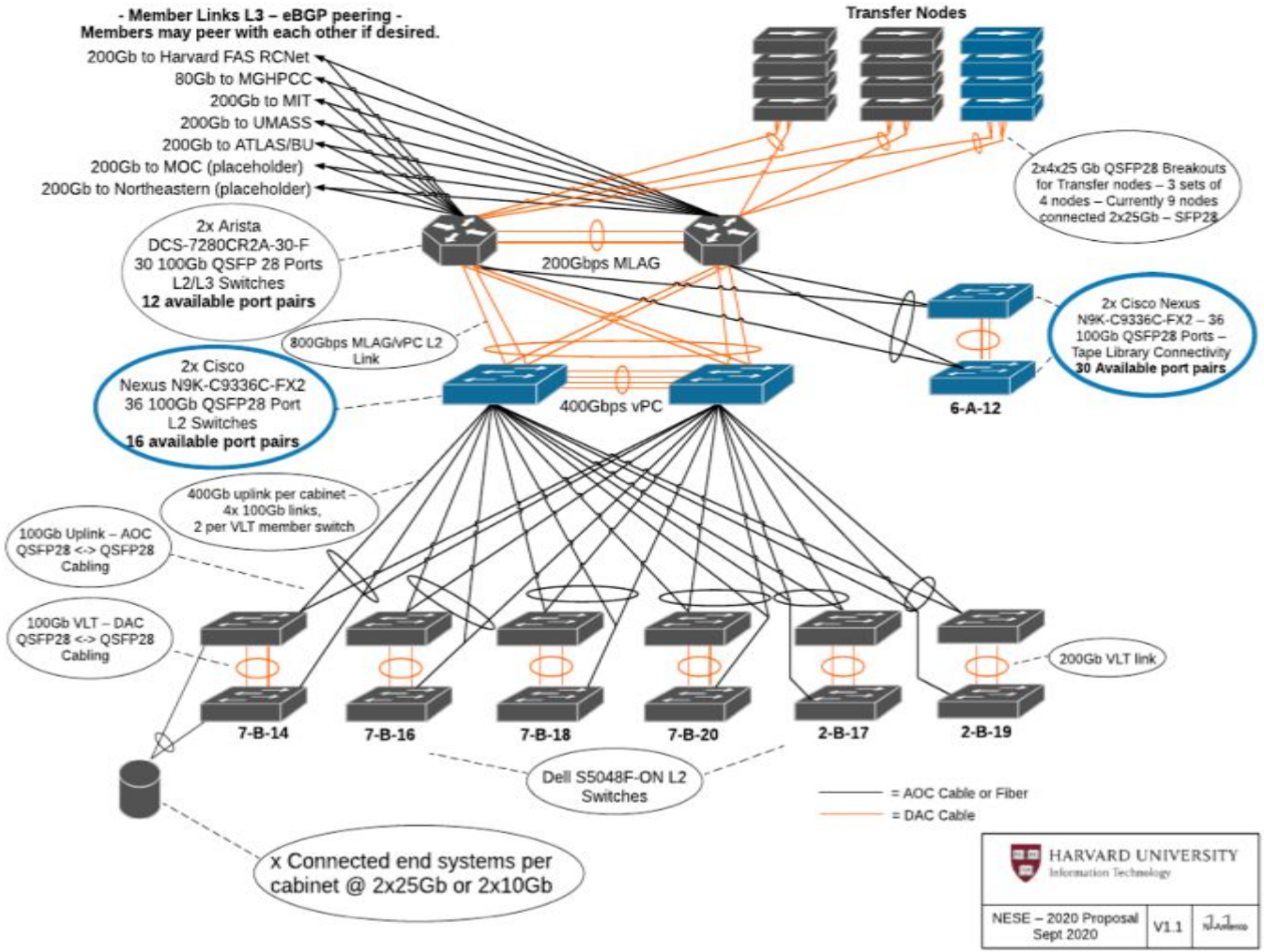
Conclusions

- It is extremely important that US-ATLAS conveys the message to BU that the most important thing for them to do is to transfer the equipment to UMass, and not to operate the legacy equipment.
  - Unfortunately, this is not the message that has been conveyed recently and it has caused delays in our plans that can compromise the first milestone.
- Progress is being made on several fronts for NET2.1
  - Team in place
  - Plan for the equipment transfer coming together
- Several open issues and external dependencies
  - Continues to be a challenge to actually get the equipment from BU
  - Network connection to ESnet in a cost-effective way
  - Storage solution - requires different mode of operation for NESE
- Many opportunities
  - Rebuilding from scratch so can explore new architectures
  - Potential for collaborations with MIT & Harvard and interesting opportunities for the future for US-ATLAS
- **Thank you very much to everyone in the US ATLAS Facility for all your help and support!**

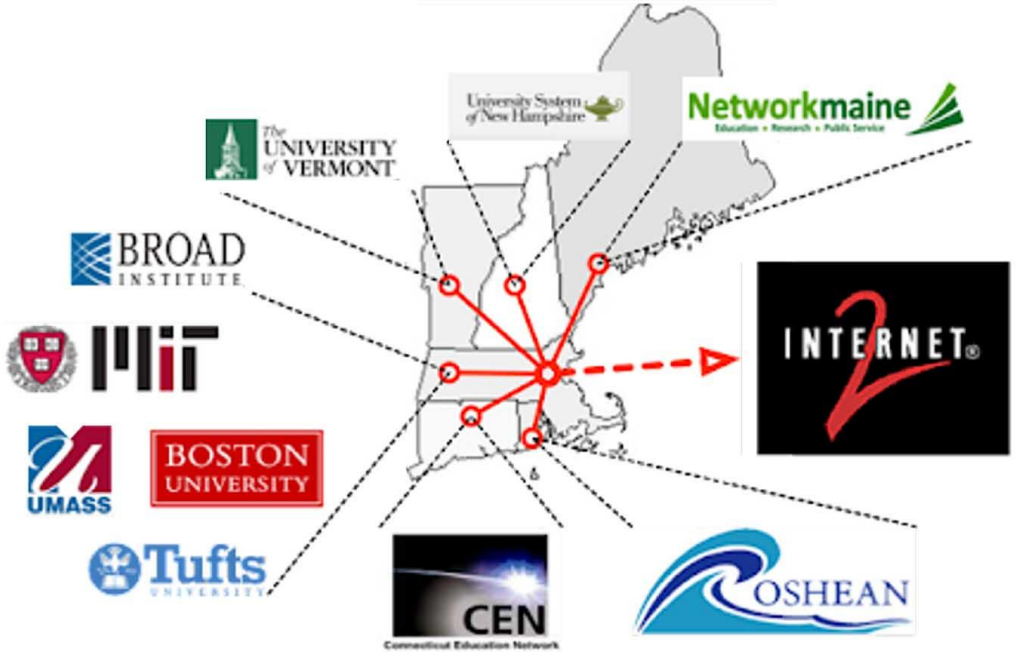
Backup

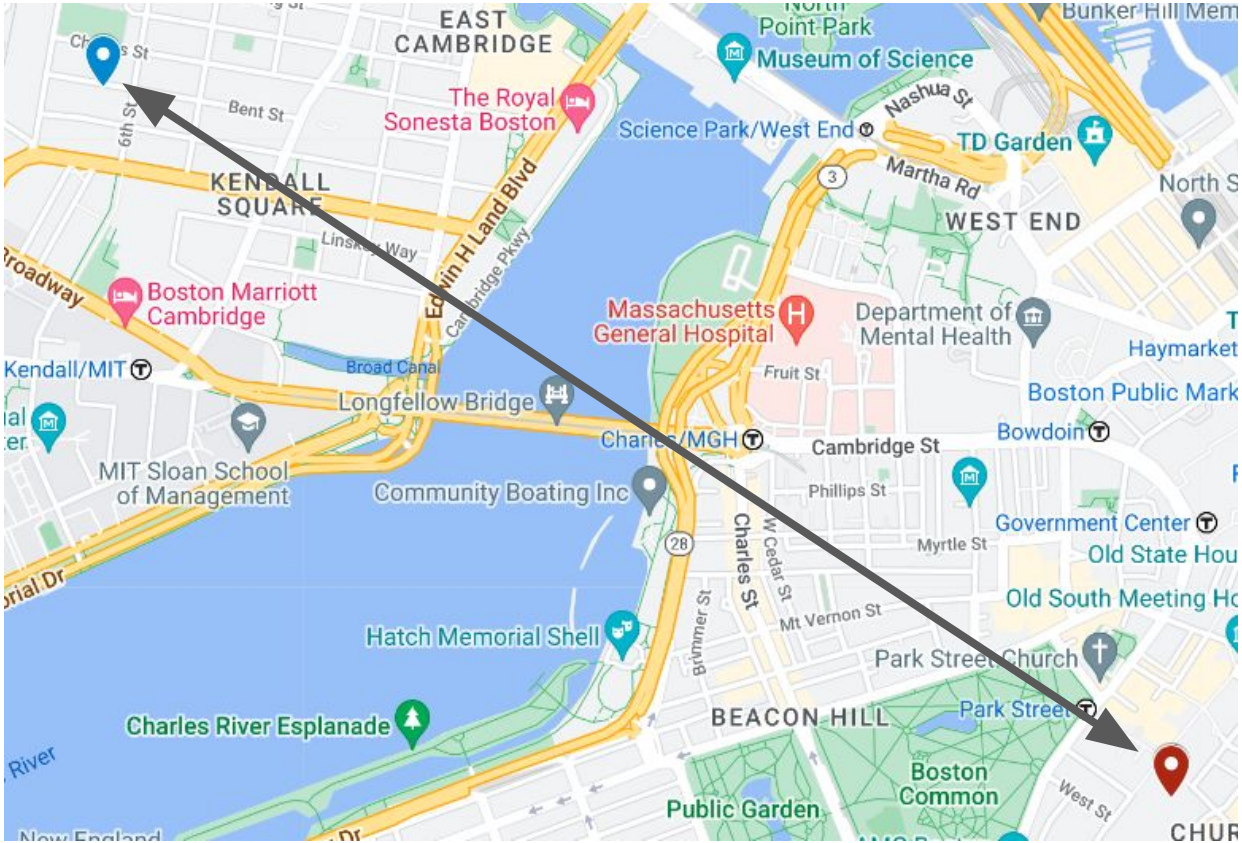


MGHPCC Pod with InRow Coolers and Hot Aisle Containment


- Member Links L3 – eBGP peering -
Members may peer with each other if desired.
200Gb to Harvard FAS RCNet
80Gb to MGHPCC
200Gb to MIT
200Gb to UMASS
200Gb to ATLAS/BU
200Gb to MOC (placeholder)
200Gb to Northeastern (placeholder)
Transfer Nodes
2x4x25 Gb QSFP28 Breakouts for Transfer nodes – 3 sets of 4 nodes – Currently 9 nodes connected 2x25Gb – SFP28
2x Arista DCS-7280CR2A-30-F 30 100Gb QSFP 28 Ports L2/L3 Switches
12 available port pairs
200Gbps MLAG
800Gbps MLAG/vPC L2 Link
2x Cisco Nexus N9K-C9336C-FX2 – 36 100Gb QSFP28 Ports – Tape Library Connectivity
30 Available port pairs
2x Cisco Nexus N9K-C9336C-FX2 36 100Gb QSFP28 Port L2 Switches
16 available port pairs
400Gbps vPC
6-A-12
400Gb uplink per cabinet – 4x 100Gb links, 2 per VLT member switch
100Gb Uplink – AOC QSFP28 <-> QSFP28 Cabling
100Gb VLT – DAC QSFP28 <-> QSFP28 Cabling
200Gb VLT link
7-B-14
7-B-16
7-B-18
7-B-20
2-B-17
2-B-19
Dell S5048F-ON L2 Switches
= AOC Cable or Fiber
= DAC Cable
x Connected end systems per cabinet @ 2x25Gb or 2x10Gb
HARVARD UNIVERSITY
Information Technology
NESE – 2020 Proposal
Sept 2020
V1.1

Northern Crossroads

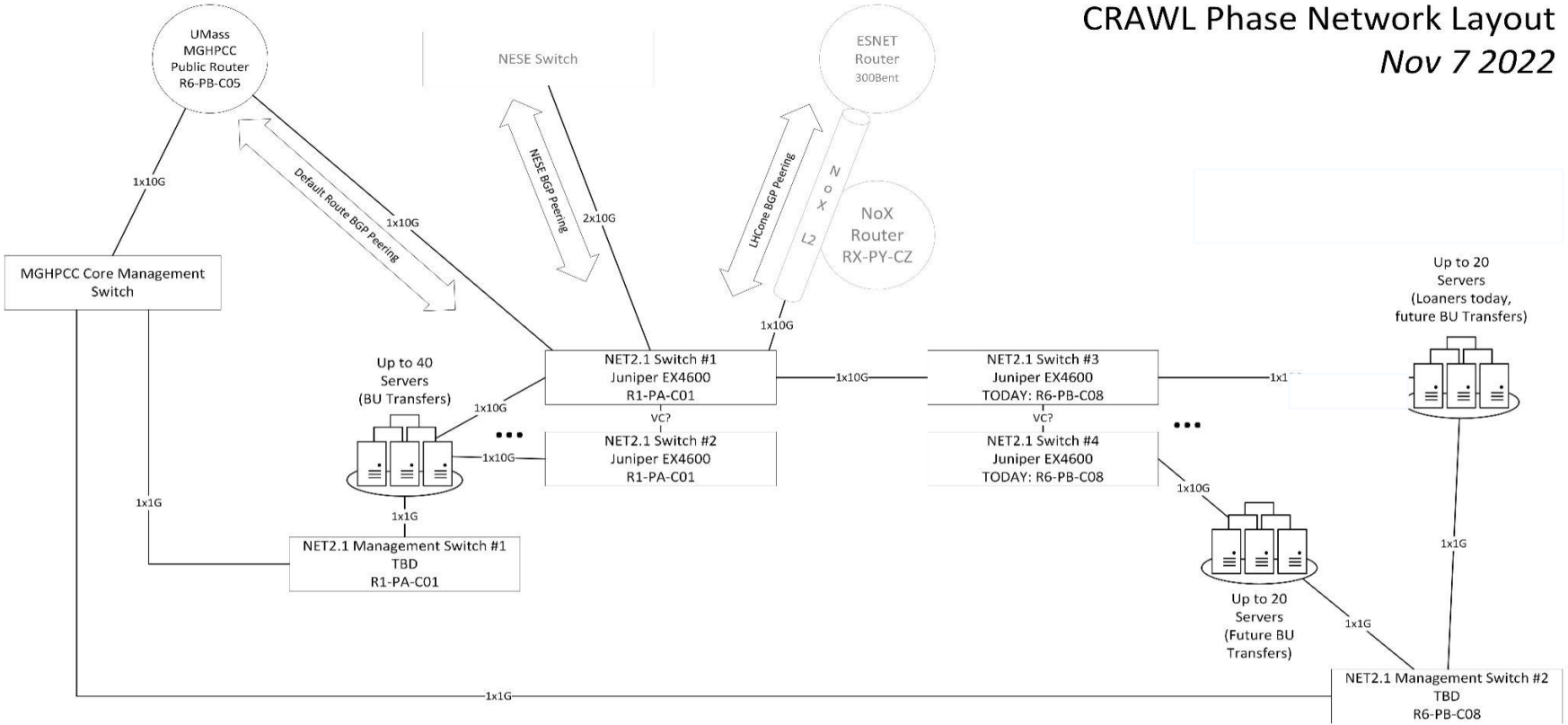


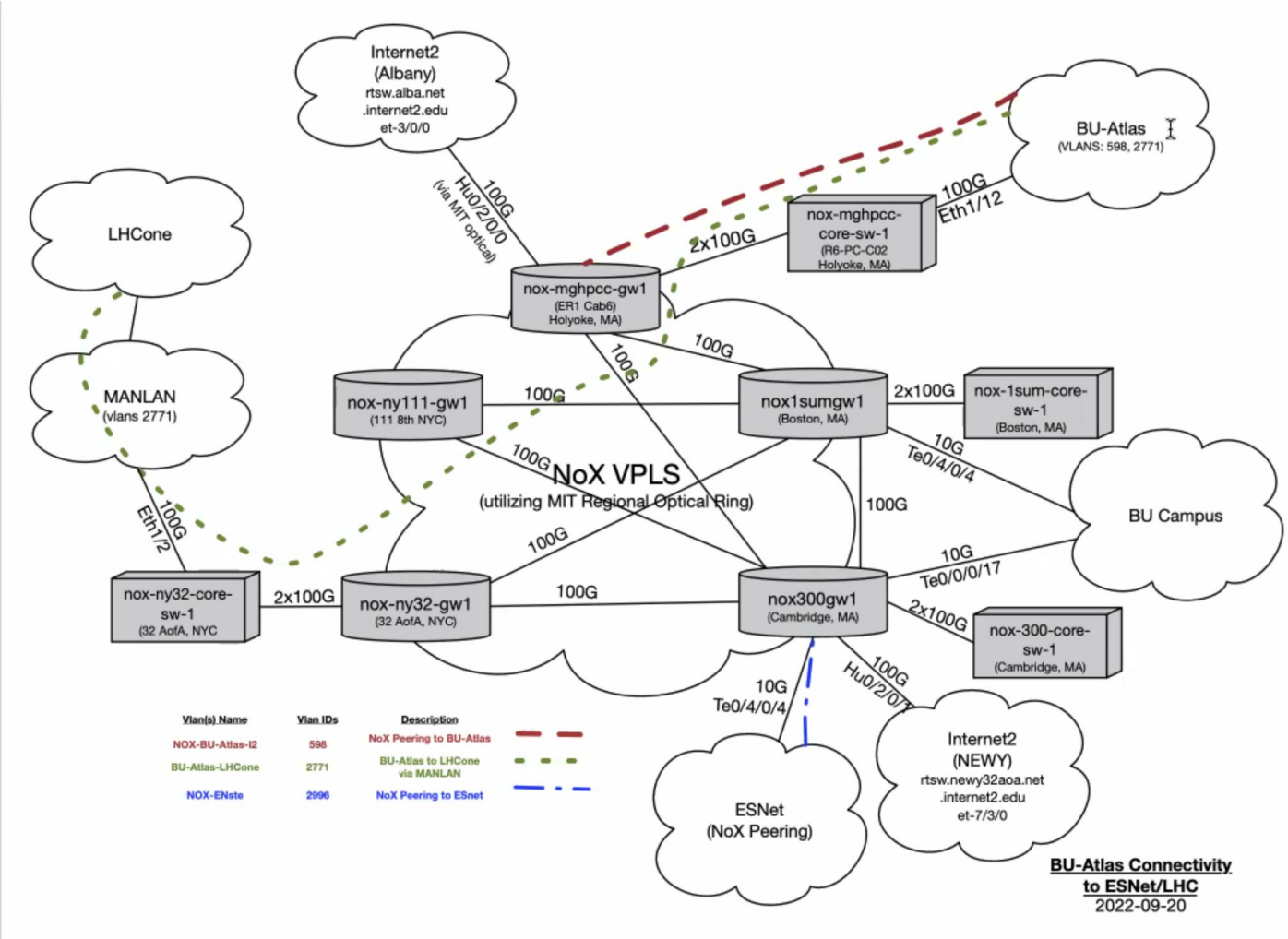

EAST CAMBRIDGE
The Royal Sonesta Boston
Point Park
Museum of Science
Bunker Hill Mem
Science Park/West End
Nashua St
TD Garden
Martha Rd
KENDALL SQUARE
WEST END
North S
Broadway
Boston Marriott Cambridge
Edwin H Land Blvd
Massachusetts General Hospital
Department of Mental Health
Haymarket
Kendall/MIT
Broad Canal
Fruit St
Longfellow Bridge
Boston Public Mark
Charles/MGH
Bowdoin
Cambridge St
MIT Sloan School of Management
Community Boating Inc
Phillips St
Government Center
Myrtle St
Old State Hou
Old South Meeting Ho
Hatch Memorial Shell
Mt Vernon St
Park Street Church
Charles River Esplanade
BEACON HILL
Park Street
Boston Common
Public Garden
West St

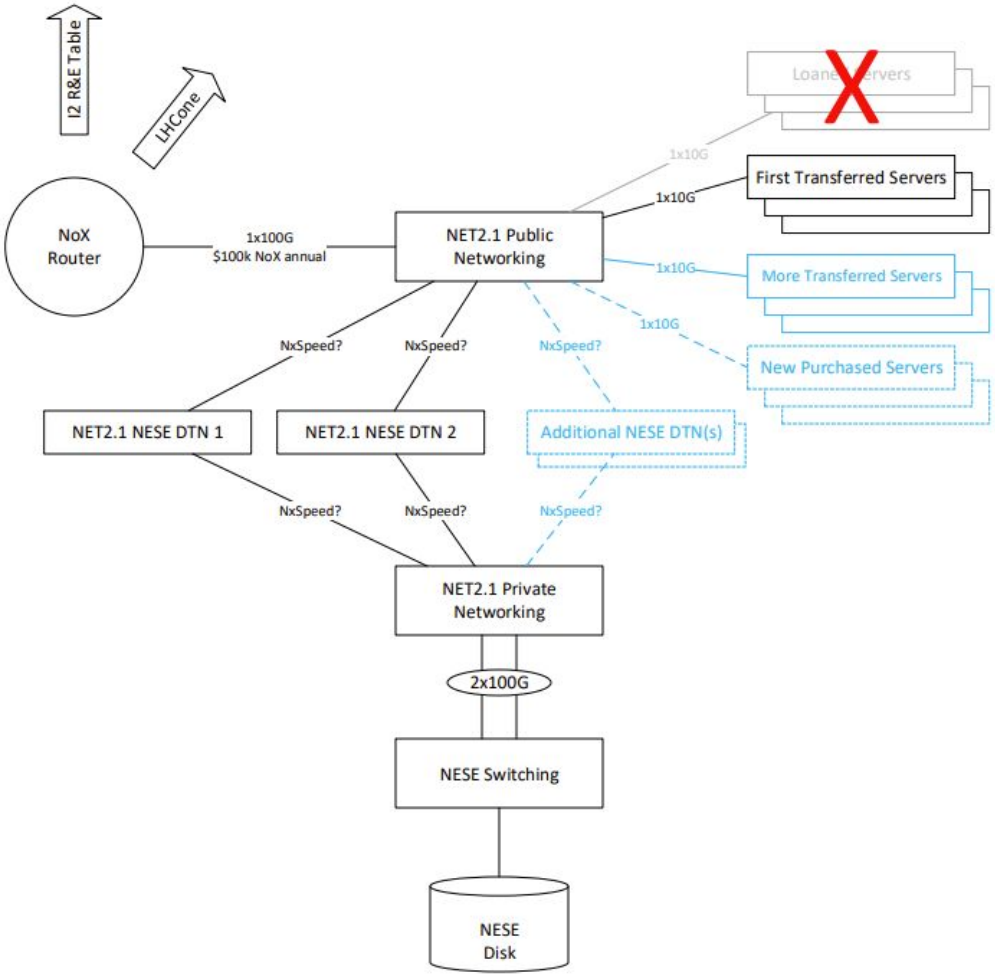
US-ATLAS NET2.1 by UMass Amherst

CRAWL Phase Network Layout

Nov 7 2022




Internet2
(Albany)
rtsw.alba.net
.internet2.edu
et-3/0/0
100G
Hu0/2/0/0
(via MIT optical)
BU-Atlas
(VLANS: 598, 2771)
100G
Eth1/12
LHCone
nox-mghpcc-
core-sw-1
(R6-PC-C02
Holyoke, MA)
2x100G
nox-mghpcc-gw1
(ER1 Cab6)
Holyoke, MA)
100G
100G
MANLAN
(vlans 2771)
nox-ny111-gw1
(111 8th NYC)
100G
nox1sumgw1
(Boston, MA)
2x100G
nox-1sum-core-
sw-1
(Boston, MA)
10G
Te0/4/0/4
100G
NoX VPLS
(utilizing MIT Regional Optical Ring)
BU Campus
100G
Eth1/2
100G
100G
10G
Te0/0/0/17
nox-ny32-core-
sw-1
(32 AofA, NYC
2x100G
nox-ny32-gw1
(32 AofA, NYC)
100G
nox300gw1
(Cambridge, MA)
2x100G
nox-300-core-
sw-1
(Cambridge, MA)
10G
Te0/4/0/4
100G
Hu0/2/0/1
Vlan(s) Name
Vlan IDs
Description
NOX-BU-Atlas-I2
598
NoX Peering to BU-Atlas
BU-Atlas-LHCone
2771
BU-Atlas to LHCone
via MANLAN
NOX-ENste
2996
NoX Peering to ESnet
ESNet
(NoX Peering)
Internet2
(NEWY)
rtsw.newy32aoa.net
.internet2.edu
et-7/3/0
BU-Atlas Connectivity
to ESNet/LHC
2022-09-20


I2 R&E Table
LHCone
NoX Router
1x100G
$100k NoX annual
NET2.1 Public Networking
Loaner Servers
1x10G
First Transferred Servers
1x10G
1x10G
More Transferred Servers
1x10G
New Purchased Servers
NxSpeed?
NxSpeed?
NxSpeed?
NET2.1 NESE DTN 1
NET2.1 NESE DTN 2
Additional NESE DTN(s)
NxSpeed?
NxSpeed?
NxSpeed?
NET2.1 Private Networking
2x100G
NESE Switching
NESE Disk