Computation Institute

Using FAX to test intra-US links

Ilija Vukotic

on behalf of the atlas-adc-federated-xrootd working group

Computing Integration and Operations meeting

February 19, 2014

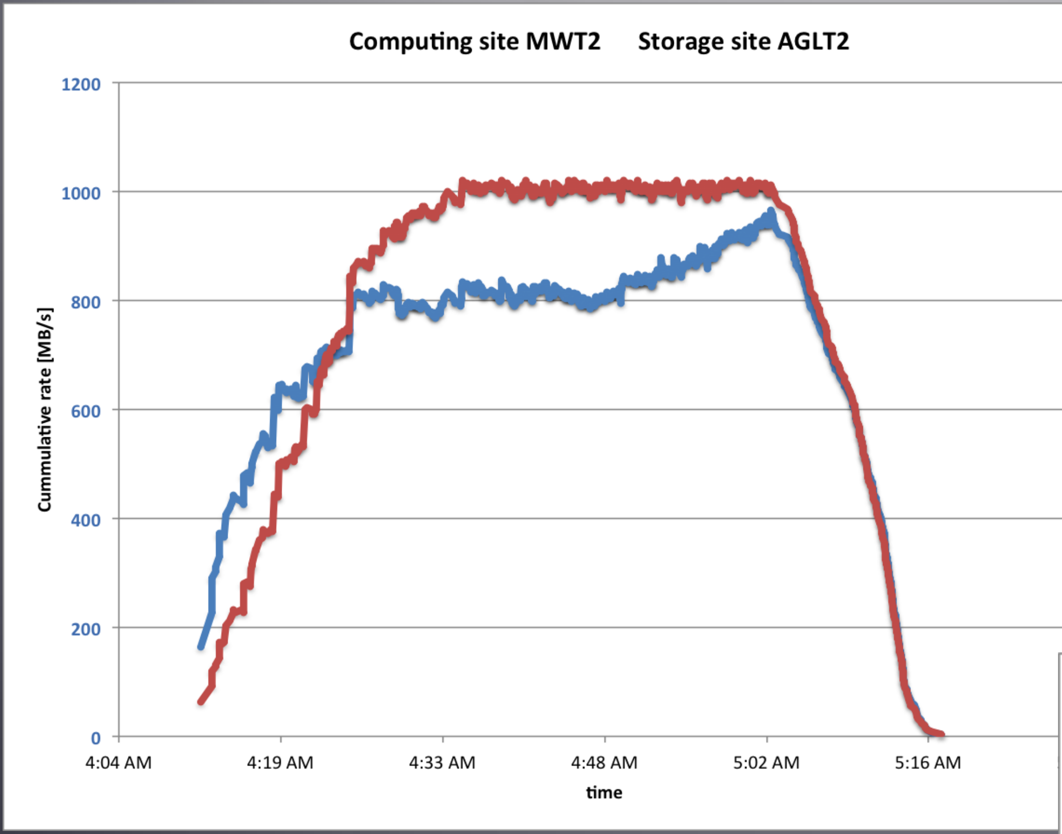
Idea

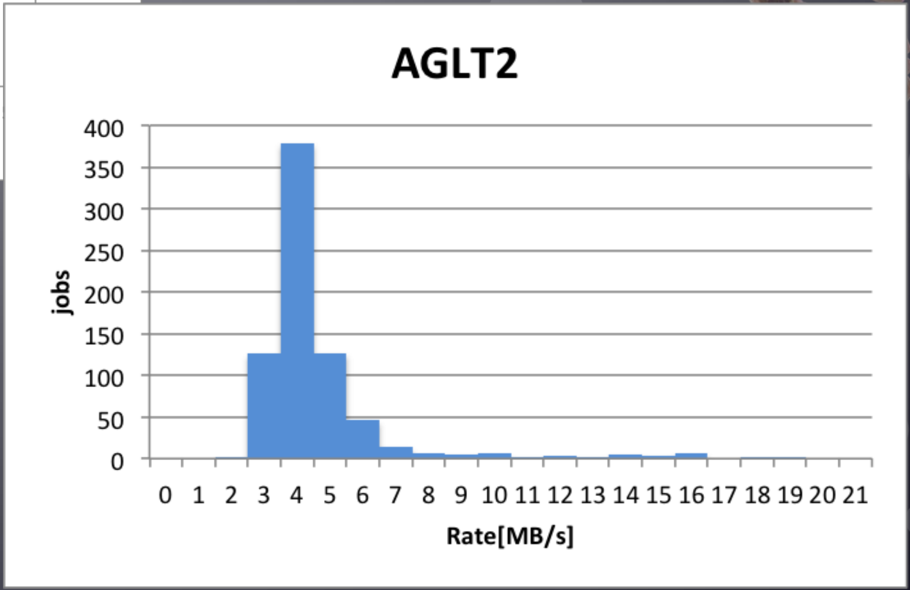
- US FAX infrastructure
  - All running with Rucio enabled N2N
  - All ready for tests
- Try to check connectivity between sites using IO intensive jobs
- Check realistic physics analysis bandwidth needs

Connectivity tests

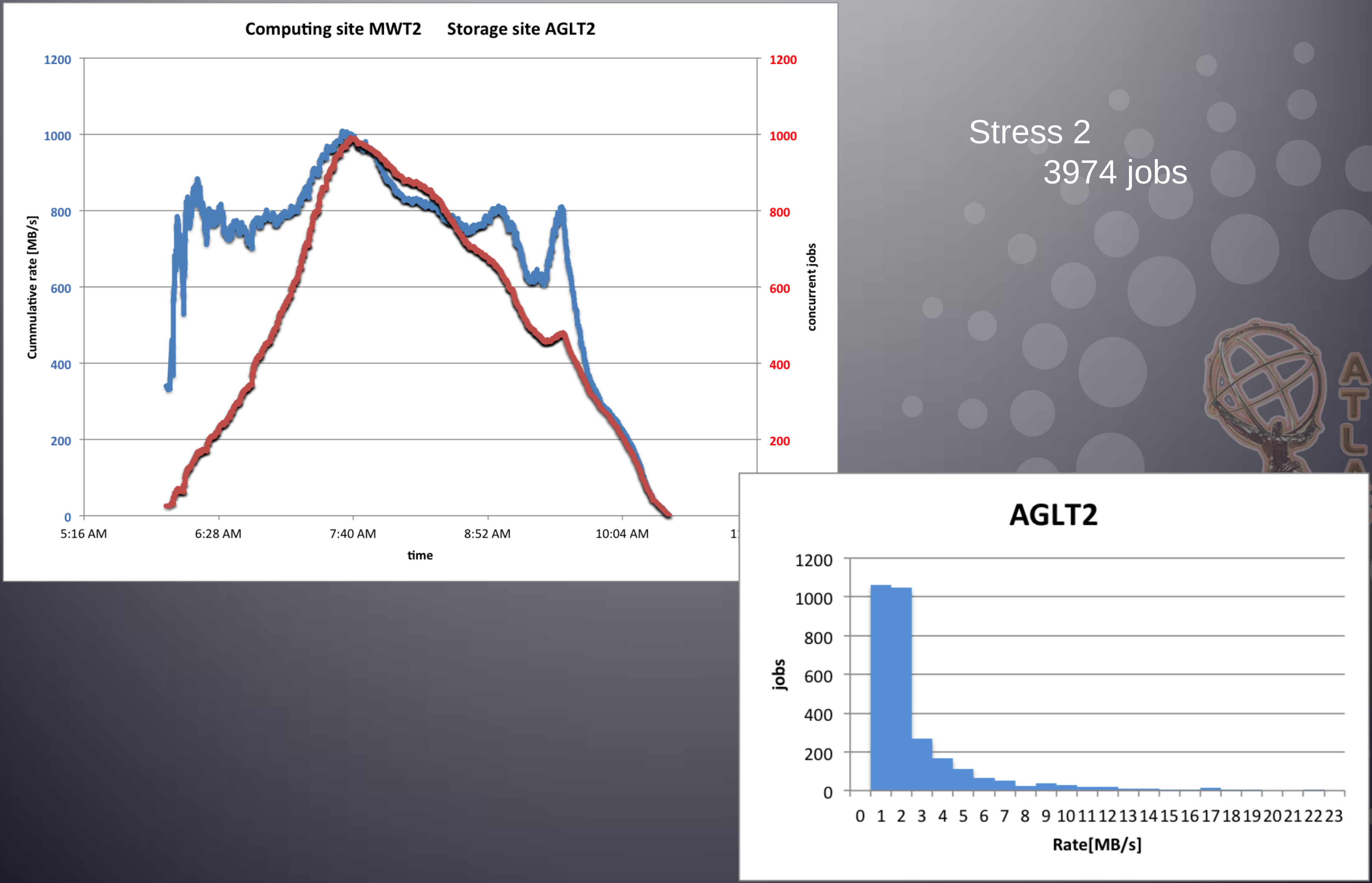
- Test:
  - Using site specific 744 FDR files (2.7 TB)
  - Reading 10% events using 30MB TTC
  - Jobs submitted using ATLAS connect
  - Running at MWT2 UC only
  - One job one file

Connectivity tests – results – AGLT2 Stress 1



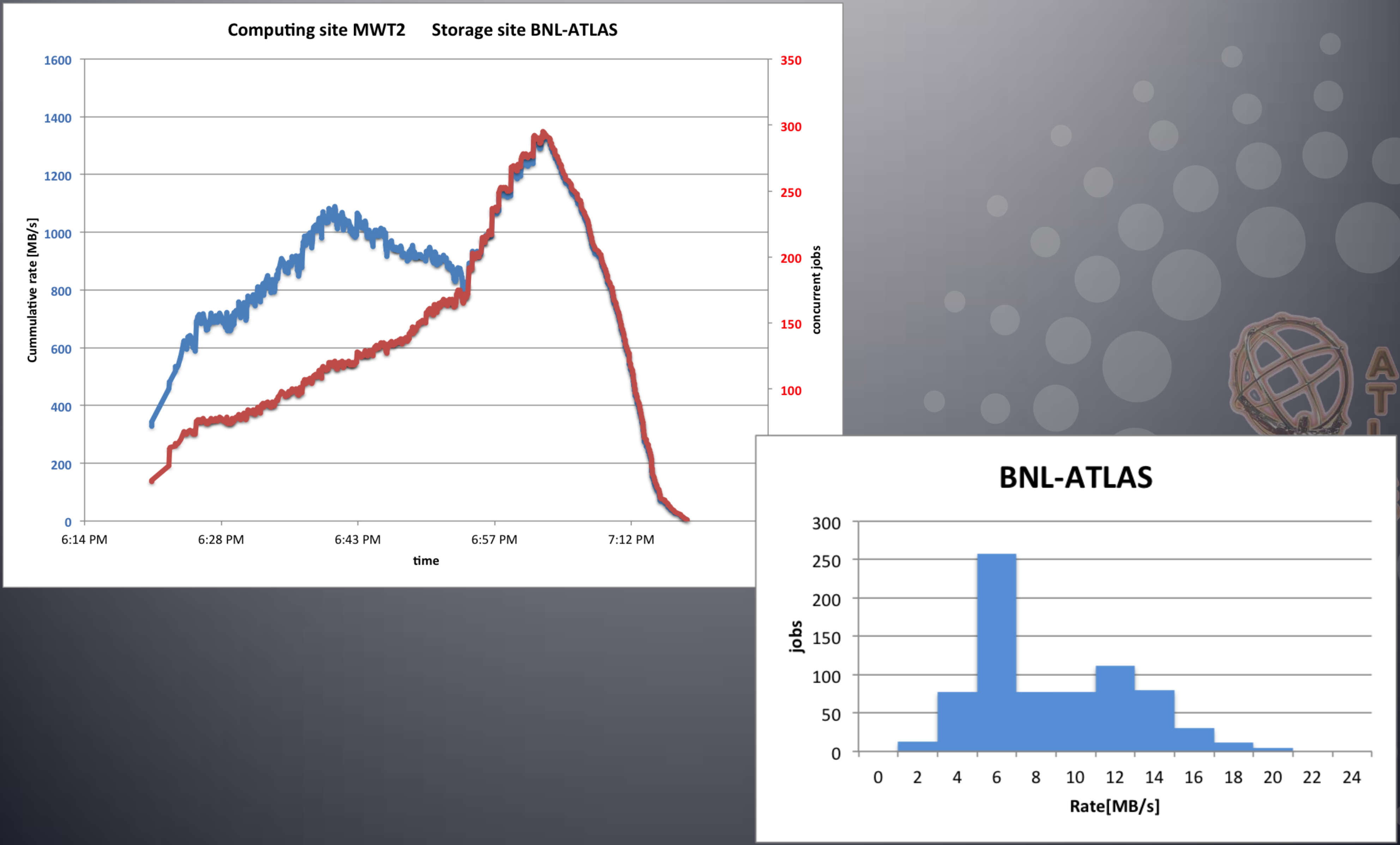


Connectivity tests – results – AGLT2 Stress 2



Stress 2
3974 jobs

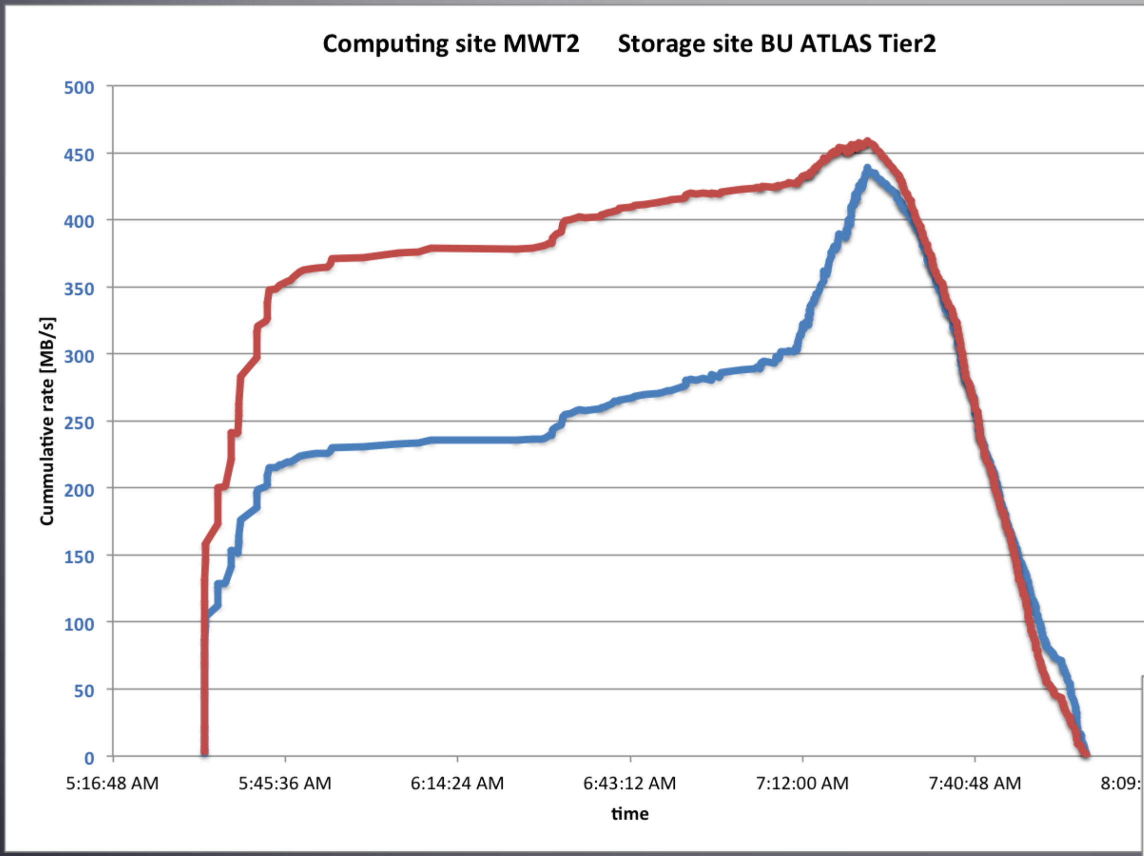
Connectivity tests – results – BNL

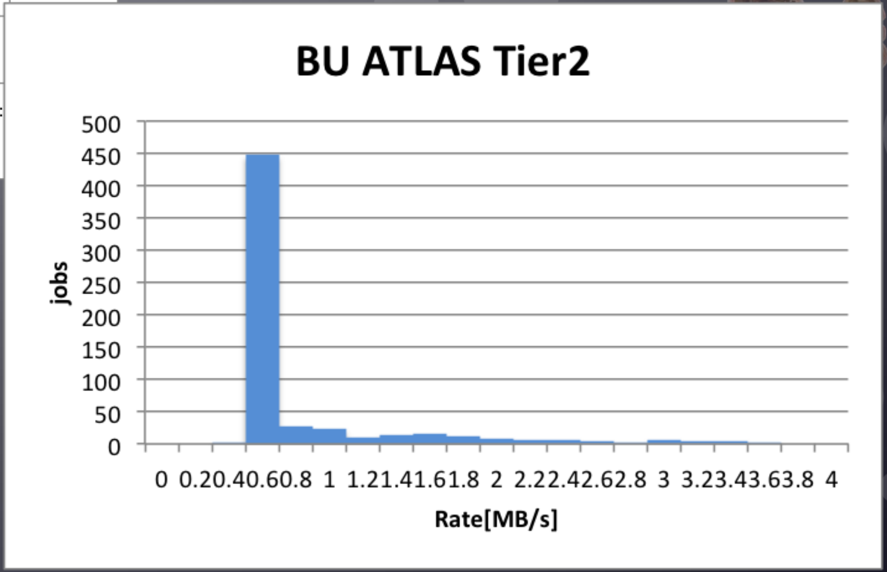


Connectivity tests – results – BNL

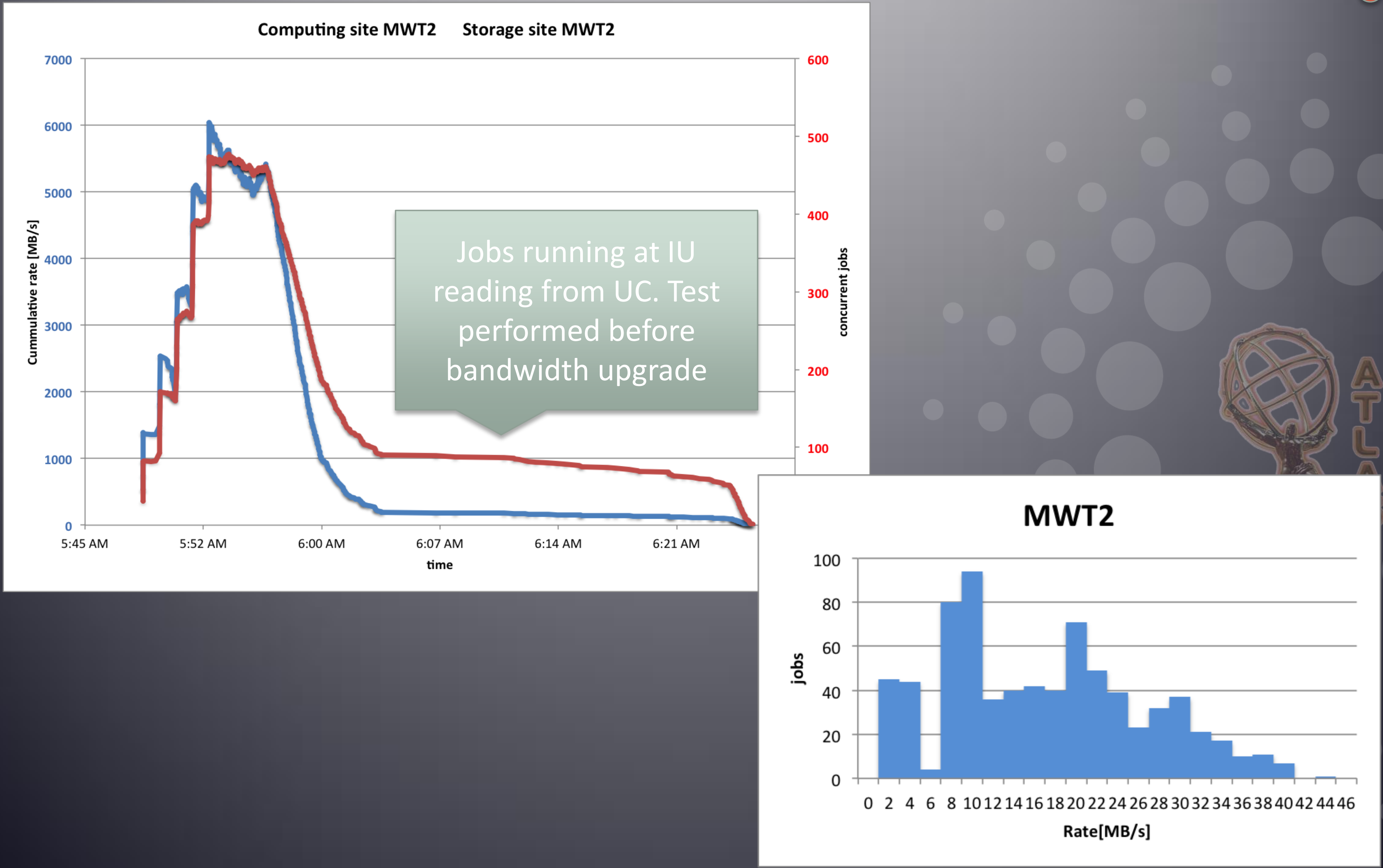
- We repeatedly tried to obtain higher bandwidth from BNL
- Used resources of MTW (UC, IU) + AGLT2 + Fresno
- In the process discovered and fixed several FAX endpoint configuration issues
- Found that BNL connection has 10Gbps limit
- Will repeat the test when 100Gbps gets established

Connectivity tests – results – BU



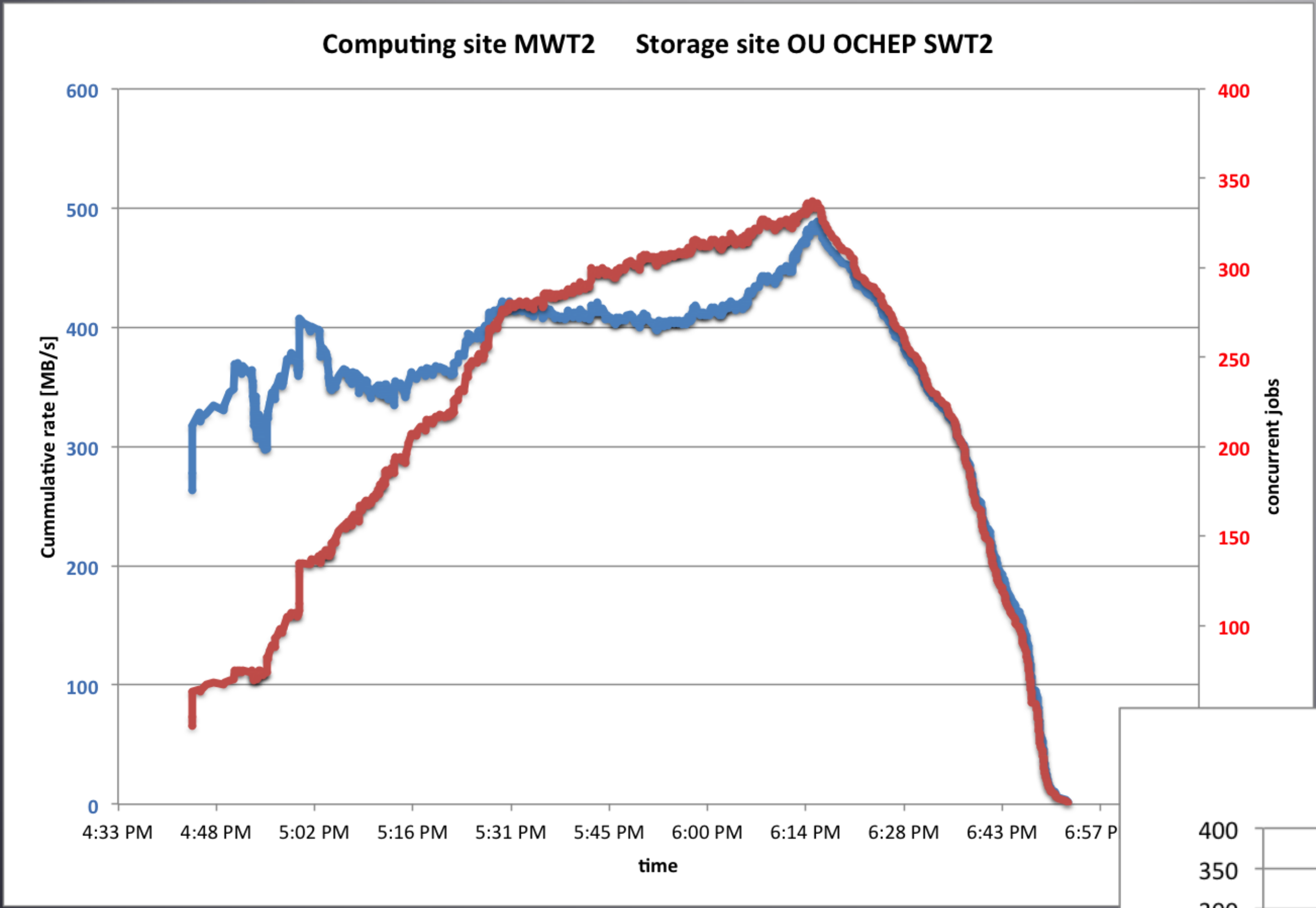


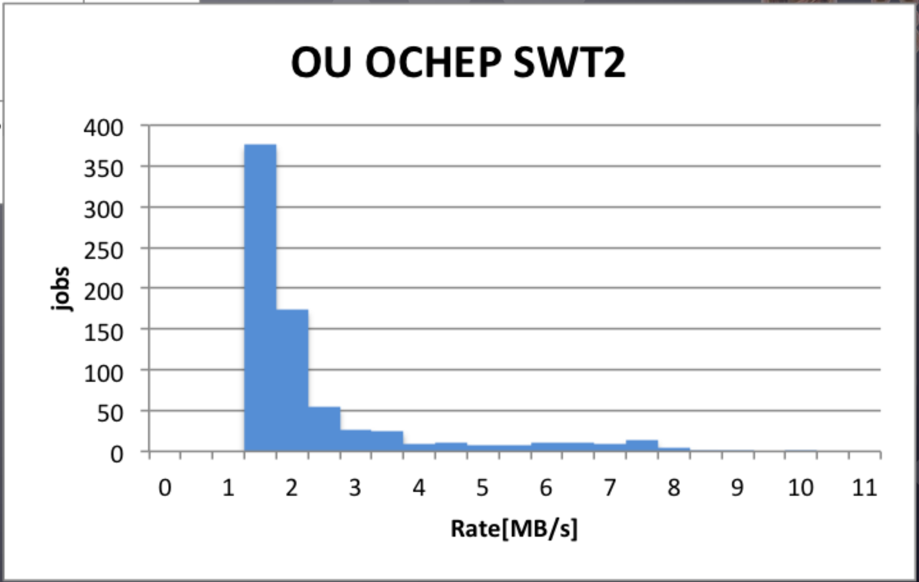
Connectivity tests – results – MWT2



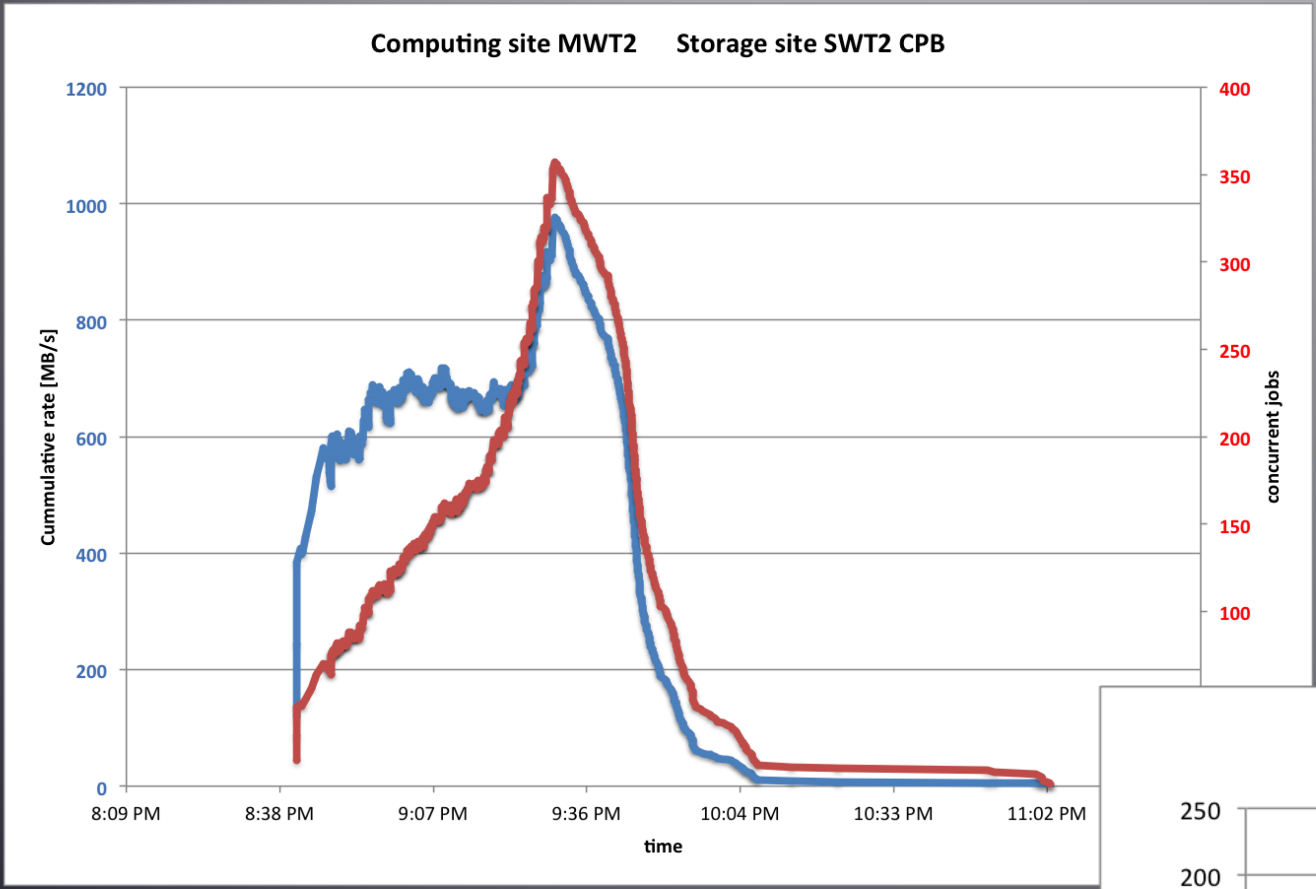
Jobs running at IU reading from UC. Test performed before bandwidth upgrade

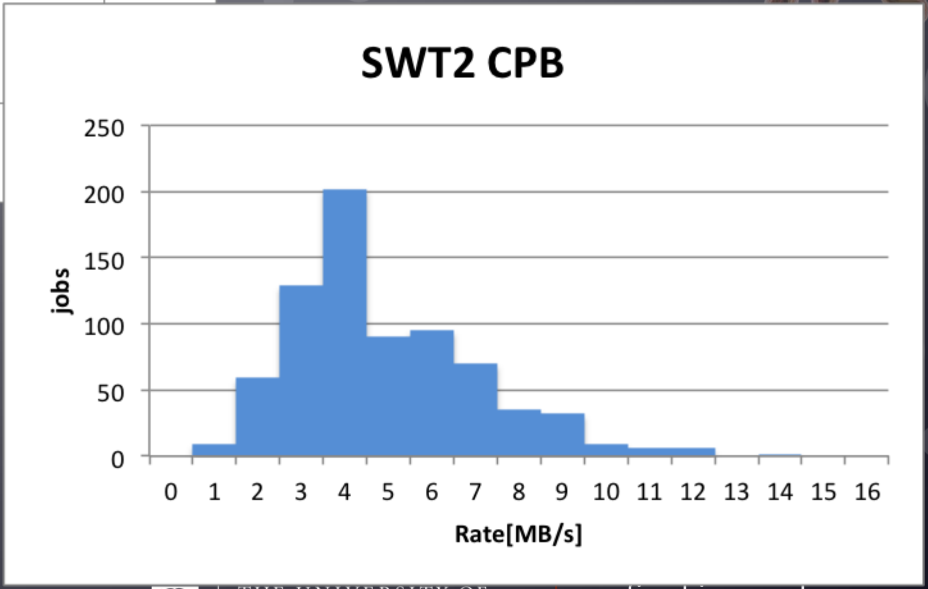
Connectivity tests – results – OU_OCHEP_SWT2



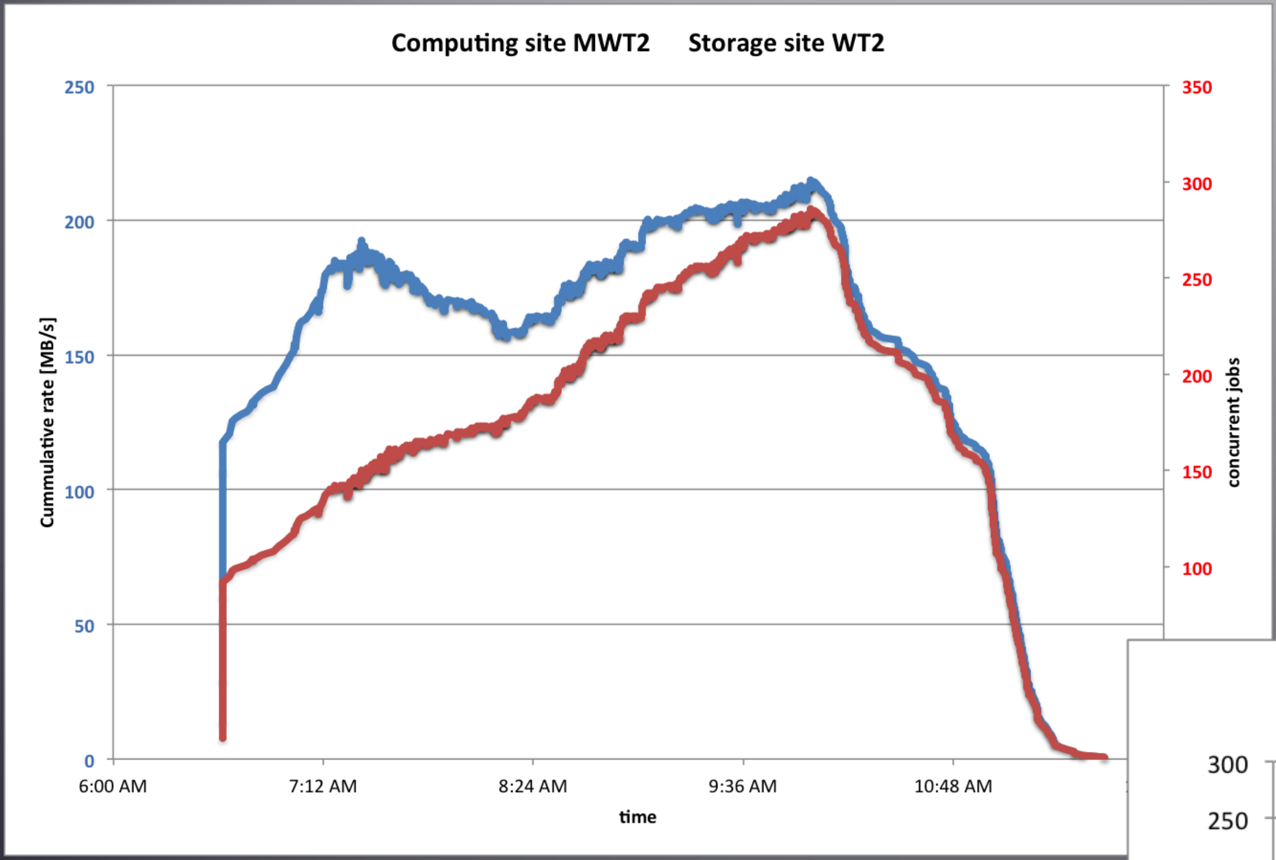


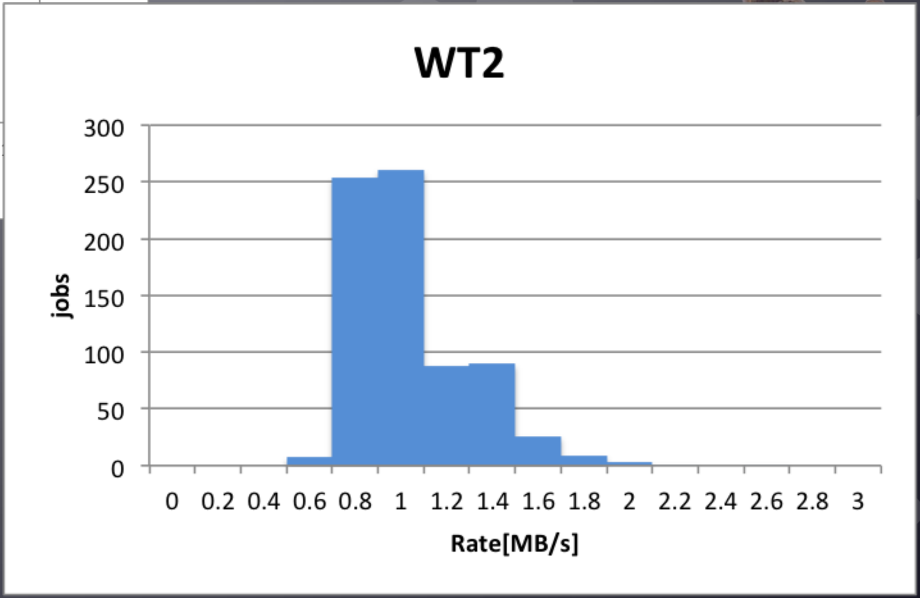
Connectivity tests – results – SWT2_CPB



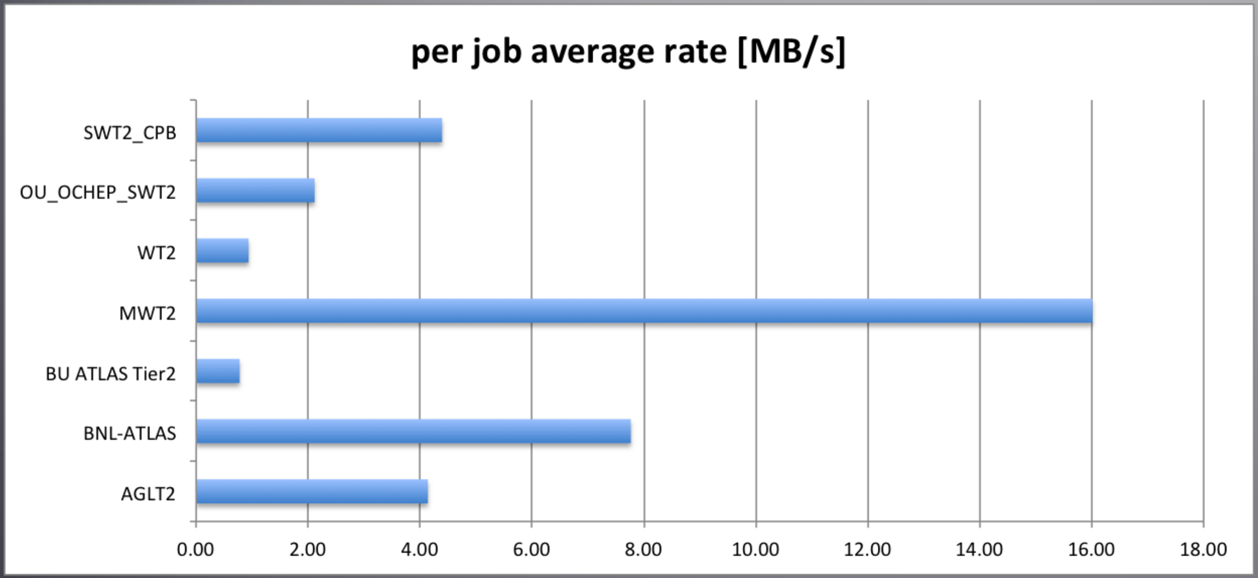


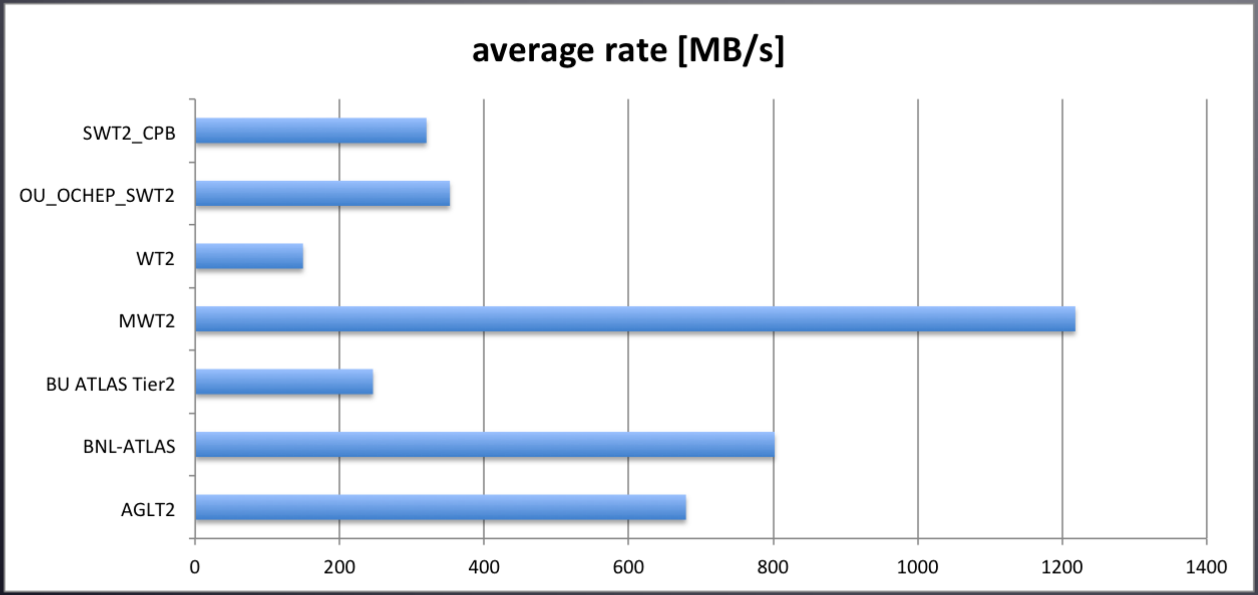
Connectivity tests – results – WT2





Connectivity tests - summary





- Very low failure rate from all endpoints
- Surprisingly small bandwidth from BU and WT2

Realistic analysis tests

- Friedrich’s Higgs to WW analysis
  - Using RootCore framework
  - All corrections applied
  - Reads 512 from 8543 branches (13% of total size)
  - Writes out <1% events and a number of histograms/TTrees
  - 10MB TTC (suboptimal)
  - Default learning phase (100 events - suboptimal)

Realistic analysis tests – results

- Having the data file cached in memory
  - 1300 ev/s
  - 20.2 MB/s of useful data
- Having the data on a local disk – single spindle, no other activity
  - 600 ev/s
  - 8.6 MB/s of useful data
  - CPU efficiency 58%

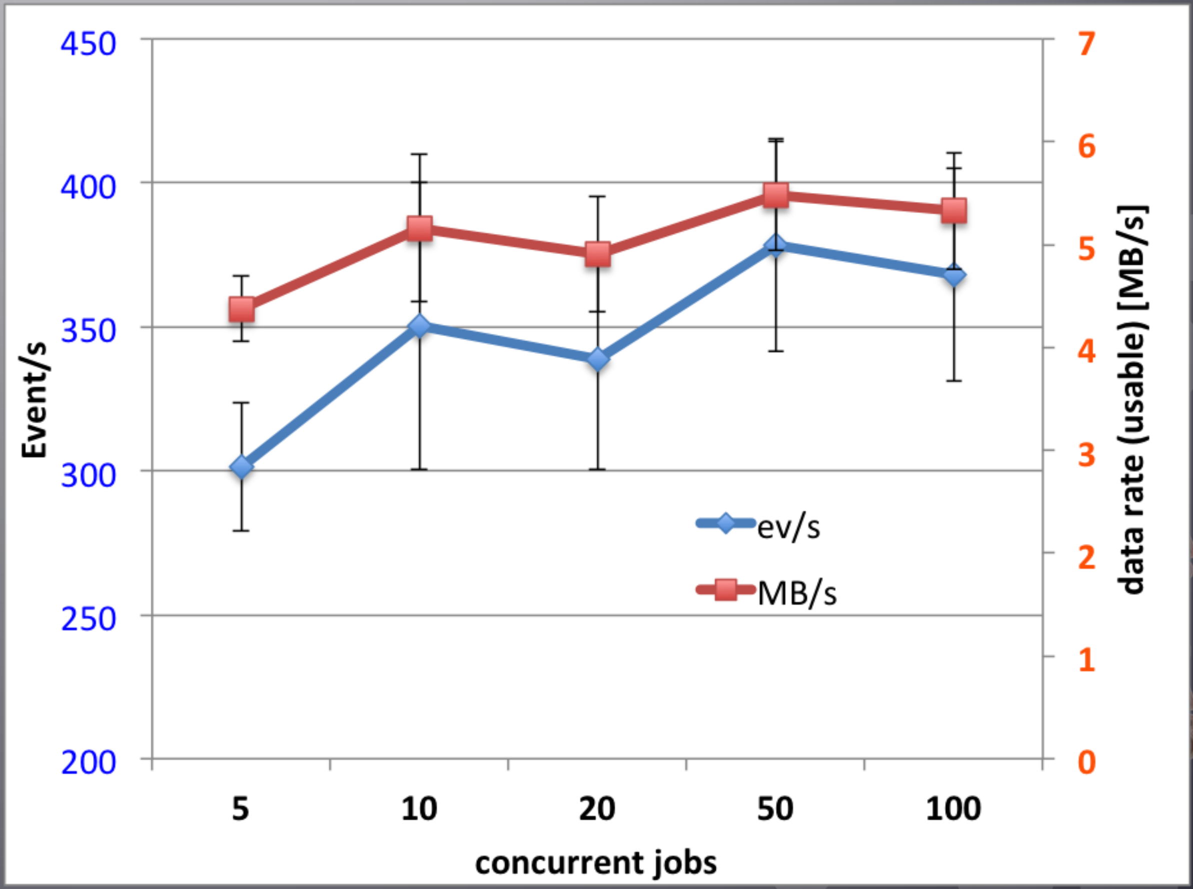
Realistic analysis tests - results

WAN access

- run at IU reading data from UC (5ms RTT)
- Each job analyzing:
  - 25 files
  - 90.5 GB
  - 832 kEvents
- Slowly ramping up in order to see when we hit a bottleneck

Realistic analysis tests - results

- Basically flat up to 500-600 MB/s
- Much larger scale needed to saturate the link now that we have confirmed 60 Gbps .
- Not much slower than local disk
- With default TTC and shorter learning phase should approach 10MB/s. Will test.



concurrent jobs	ev/s	err ev/s	MB/s	err MB/s
5	301	22	4.4	0.3
10	350	50	5.2	0.7
20	339	39	4.9	0.6
50	378	37	5.5	0.5
100	368	37	5.3	0.6

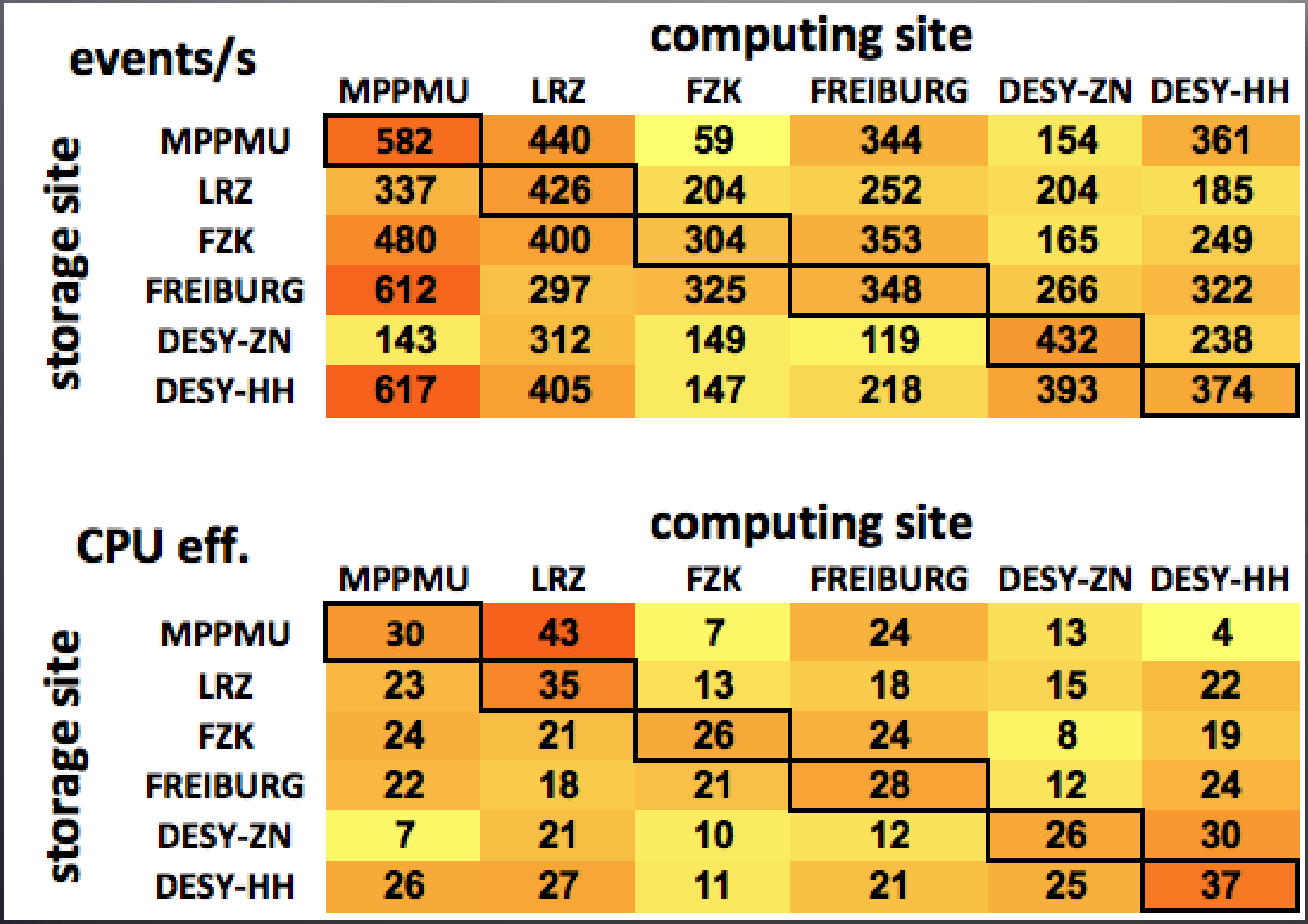
Conclusions

- FAX endpoint at US sites much more stable now.
- Stability and low failure rate make it good enough for direct access use.
- With special direct access test was able to reach the link limits of most endpoints.
- Realistic analysis used less bandwidth (5MB/s) than expected 10 MB/s.
- Will repeat the tests at larger scale with/without further optimization.
- Will try to saturate 100Gbps MWT2-BNL link when it comes online.

Reserve

WAN Testing in DE cloud

- Functional HC tests by Friedrich
- We should expect similar numbers from US

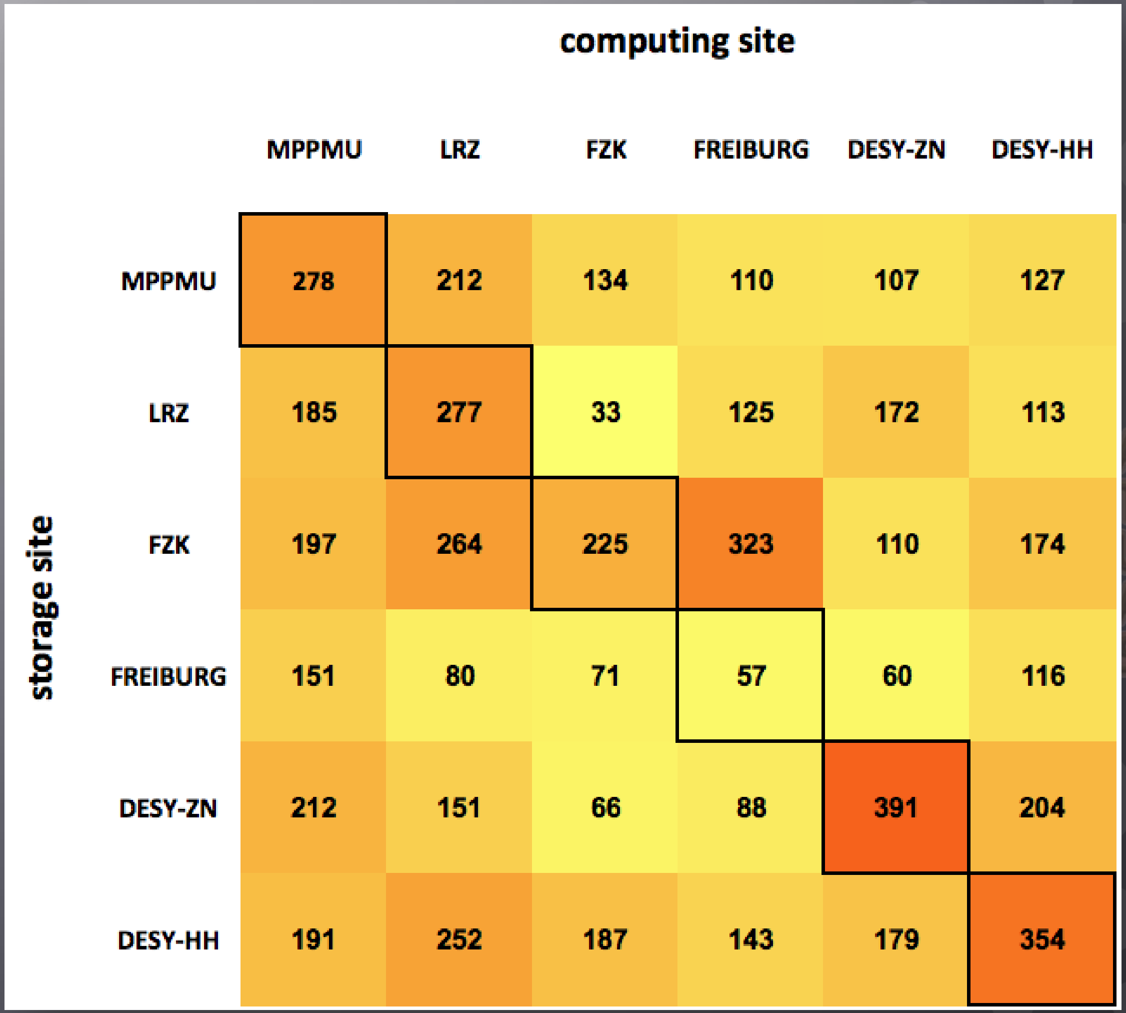


events/s	computing site → MPPMU	LRZ	FZK	FREIBURG	DESY-ZN	DESY-HH
storage site: MPPMU	582	440	59	344	154	361
LRZ	337	426	204	252	204	185
FZK	480	400	304	353	165	249
FREIBURG	612	297	325	348	266	322
DESY-ZN	143	312	149	119	432	238
DESY-HH	617	405	147	218	393	374

CPU eff.	computing site → MPPMU	LRZ	FZK	FREIBURG	DESY-ZN	DESY-HH
storage site: MPPMU	30	43	7	24	13	4
LRZ	23	35	13	18	15	22
FZK	24	21	26	24	8	19
FREIBURG	22	18	21	28	12	24
DESY-ZN	7	21	10	12	26	30
DESY-HH	26	27	11	21	25	37

WAN Load Test (200 job scale) – DE cloud

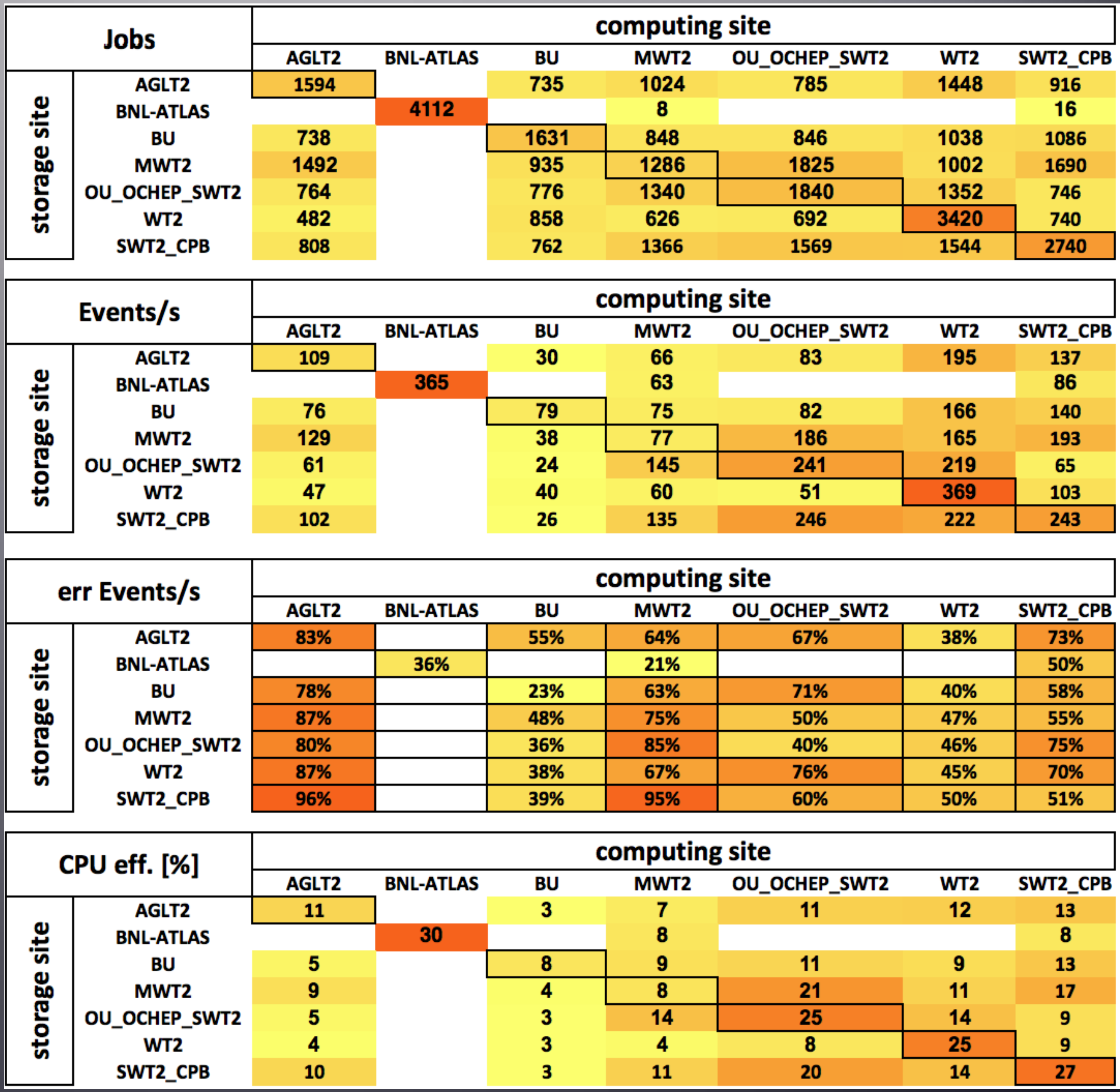
- Some uncertainty of #concurrently running jobs (not directly controllable)
- Indicates reasonable opportunity for re-brokering



		computing site					
		MPPMU	LRZ	FZK	FREIBURG	DESY-ZN	DESY-HH
storage site	MPPMU	278	212	134	110	107	127
	LRZ	185	277	33	125	172	113
	FZK	197	264	225	323	110	174
	FREIBURG	151	80	71	57	60	116
	DESY-ZN	212	151	66	88	391	204
	DESY-HH	191	252	187	143	179	354

WAN Testing in US cloud

- Numbers lower than in DE cloud.
- Possible explanations:
  - Diagonal elements – maybe we overprovision CPU at the larger scale than DE
  - Off-diagonal elements - suboptimal TTC size and learning phase result in larger than optimal RTT penalty. Size of US makes this price higher for US cloud.
  - …



Jobs		computing site						
		AGLT2	BNL-ATLAS	BU	MWT2	OU_OCHEP_SWT2	WT2	SWT2_CPB
storage site	AGLT2	1594		735	1024	785	1448	916
	BNL-ATLAS		4112		8			16
	BU	738		1631	848	846	1038	1086
	MWT2	1492		935	1286	1825	1002	1690
	OU_OCHEP_SWT2	764		776	1340	1840	1352	746
	WT2	482		858	626	692	3420	740
	SWT2_CPB	808		762	1366	1569	1544	2740

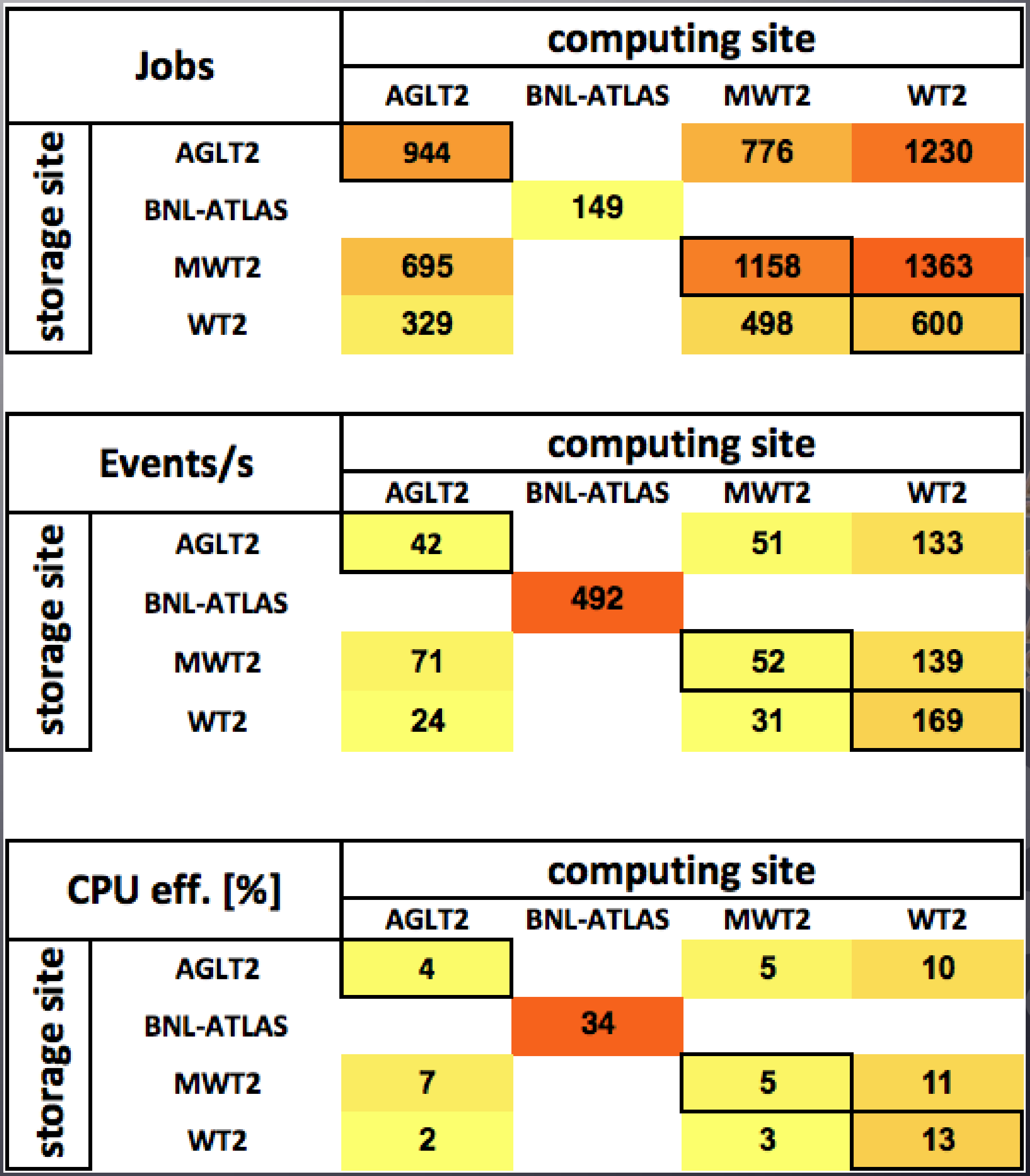
Events/s		computing site						
		AGLT2	BNL-ATLAS	BU	MWT2	OU_OCHEP_SWT2	WT2	SWT2_CPB
storage site	AGLT2	109		30	66	83	195	137
	BNL-ATLAS		365		63			86
	BU	76		79	75	82	166	140
	MWT2	129		38	77	186	165	193
	OU_OCHEP_SWT2	61		24	145	241	219	65
	WT2	47		40	60	51	369	103
	SWT2_CPB	102		26	135	246	222	243

err Events/s		computing site						
		AGLT2	BNL-ATLAS	BU	MWT2	OU_OCHEP_SWT2	WT2	SWT2_CPB
storage site	AGLT2	83%		55%	64%	67%	38%	73%
	BNL-ATLAS		36%		21%			50%
	BU	78%		23%	63%	71%	40%	58%
	MWT2	87%		48%	75%	50%	47%	55%
	OU_OCHEP_SWT2	80%		36%	85%	40%	46%	75%
	WT2	87%		38%	67%	76%	45%	70%
	SWT2_CPB	96%		39%	95%	60%	50%	51%

CPU eff. [%]		computing site						
		AGLT2	BNL-ATLAS	BU	MWT2	OU_OCHEP_SWT2	WT2	SWT2_CPB
storage site	AGLT2	11		3	7	11	12	13
	BNL-ATLAS		30		8			8
	BU	5		8	9	11	9	13
	MWT2	9		4	8	21	11	17
	OU_OCHEP_SWT2	5		3	14	25	14	9
	WT2	4		3	4	8	25	9
	SWT2_CPB	10		3	11	20	14	27

WAN Load Test (200 job scale) – US cloud

- Surprisingly low figures
- Under investigation



Jobs		computing site			
		AGLT2	BNL-ATLAS	MWT2	WT2
storage site	AGLT2	944		776	1230
	BNL-ATLAS		149		
	MWT2	695		1158	1363
	WT2	329		498	600

Events/s		computing site			
		AGLT2	BNL-ATLAS	MWT2	WT2
storage site	AGLT2	42		51	133
	BNL-ATLAS		492		
	MWT2	71		52	139
	WT2	24		31	169

CPU eff. [%]		computing site			
		AGLT2	BNL-ATLAS	MWT2	WT2
storage site	AGLT2	4		5	10
	BNL-ATLAS		34		
	MWT2	7		5	11
	WT2	2		3	13