TIFR HTCondor on Microsoft Azure

Brij Kishor Jashal - Tata Institute of Fundamental Research
Xavier Pillons - Microsoft

On behalf of everyone who worked on this

Disclaimer

“Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the TIFR, the Government of India, or any agency thereof.”

T2_IN_TIFR CMS Pledge.

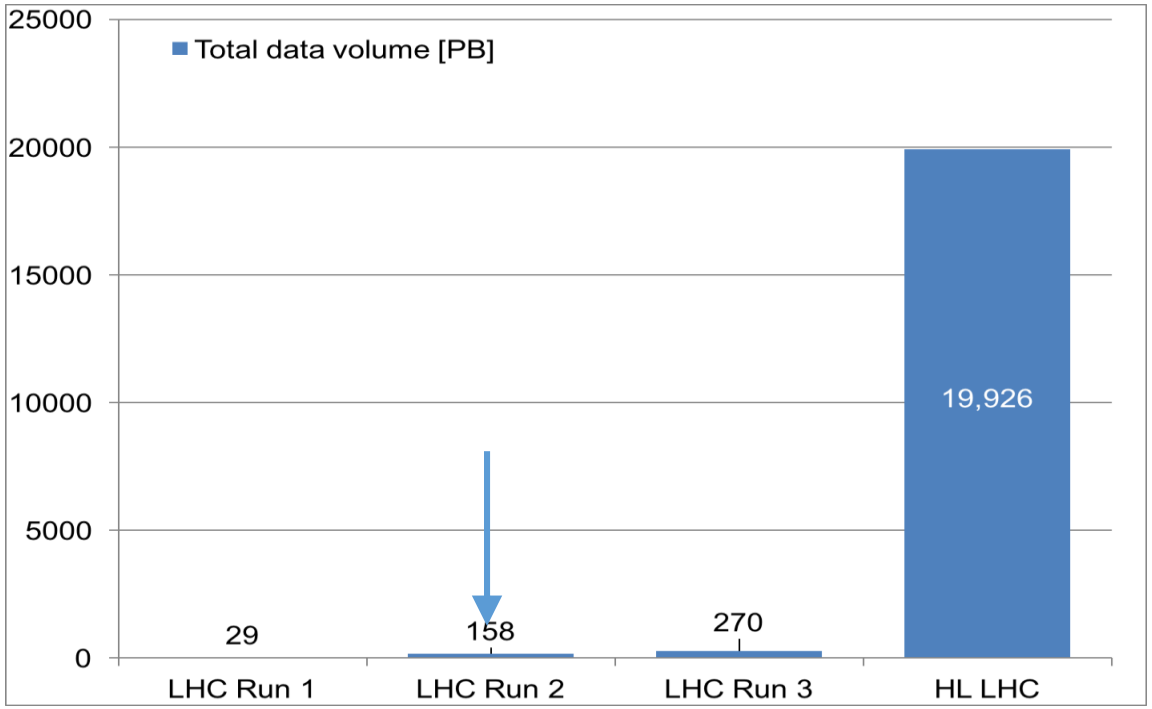
Site Name	Year	VO	Pledge Type	Resources Pledged	% of required resource
T2_IN_TIFR	2016	CMS	CPU HEPSPEC06	13,120	~1.7%
			Disk (TB)	1,980	~3%
	2017		CPU HEPSPEC06	23,000	~2.4%
			Disk (TB)	3,000	~5%

+

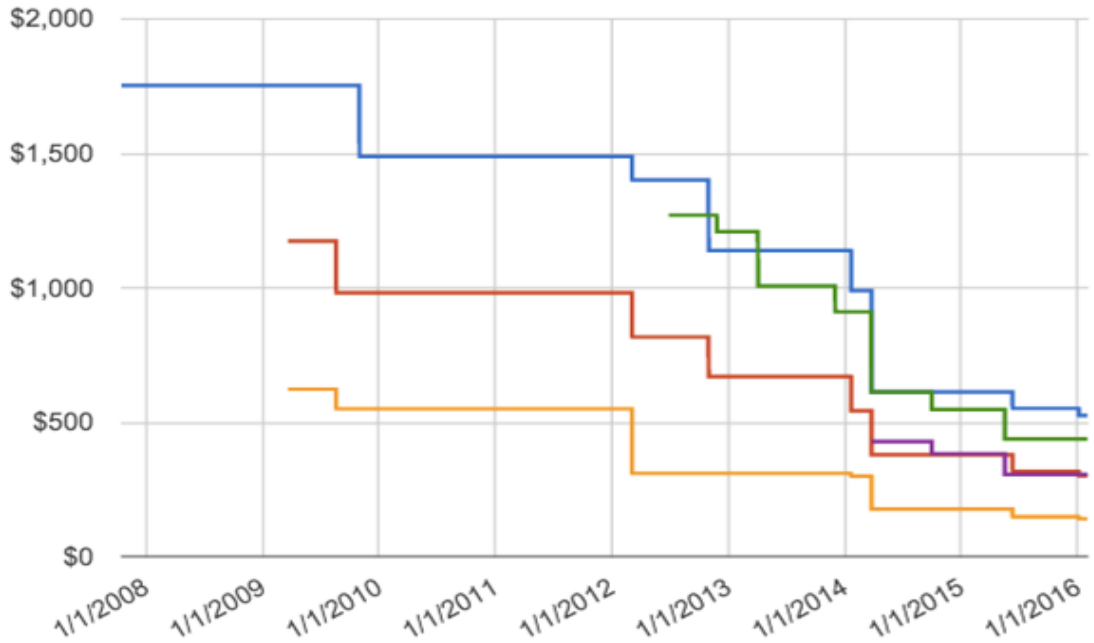
Local T3 condor pool

HEP Computing needs.

- High Energy Physics computing will need 10-100x current capacity



- Scale of industry at or above R&D
  - Commercial clouds offering increased **value** for decreased **cost** compared to the past



System to leverage global cloud infrastructure of big players.

Reference –http://research.cs.wisc.edu/htcondor/HTCondorWeek2017/presentations/WedTimm_GCE.pdf

Leveraging the global cloud presence to meet HEP computing needs becomes important.

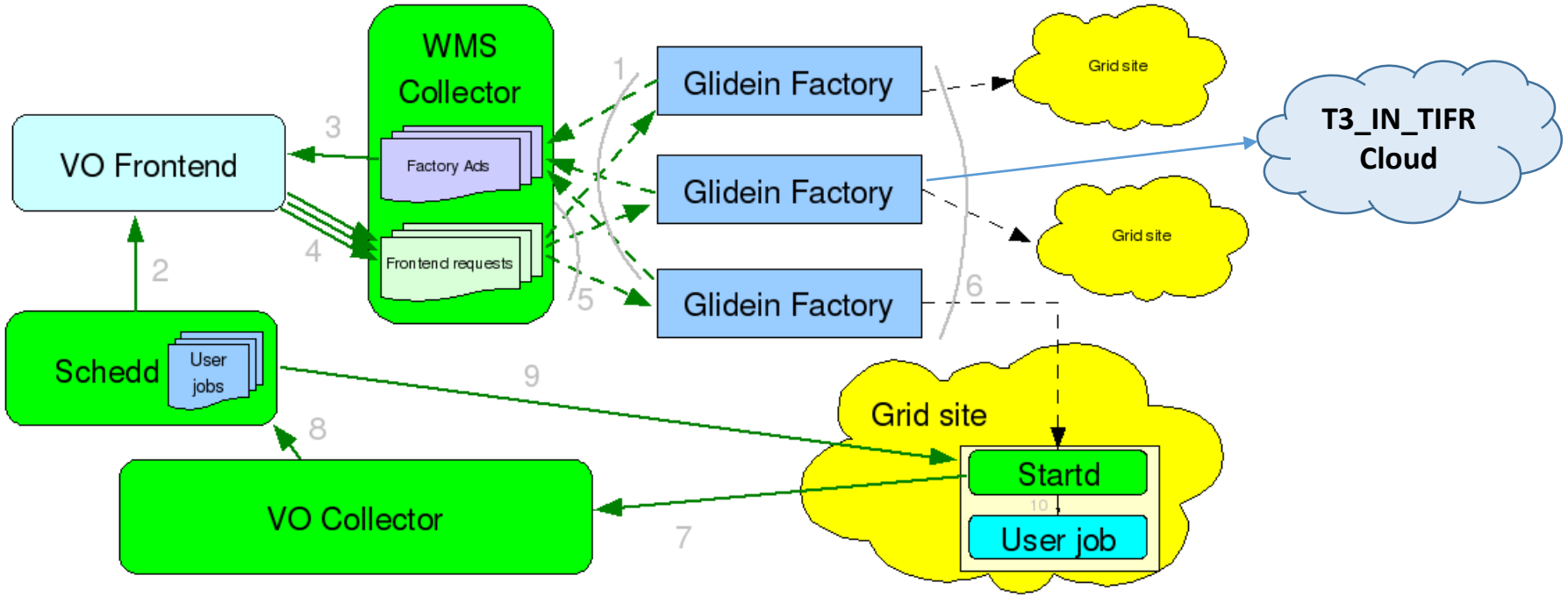
TIFR Azure Cloud

- Engagement with Microsoft started in Nov-2016
- MS Cloud Datacenters, three in India (Mumbai, Chennai, Pune)
- MS Grant in terms of resources + development resources for GAHP and Condor_Annex

Draft Deployment Scheme for TIFR Proof of Concept (200 Core Test)

Integrating Azure resources in CMS Global pool

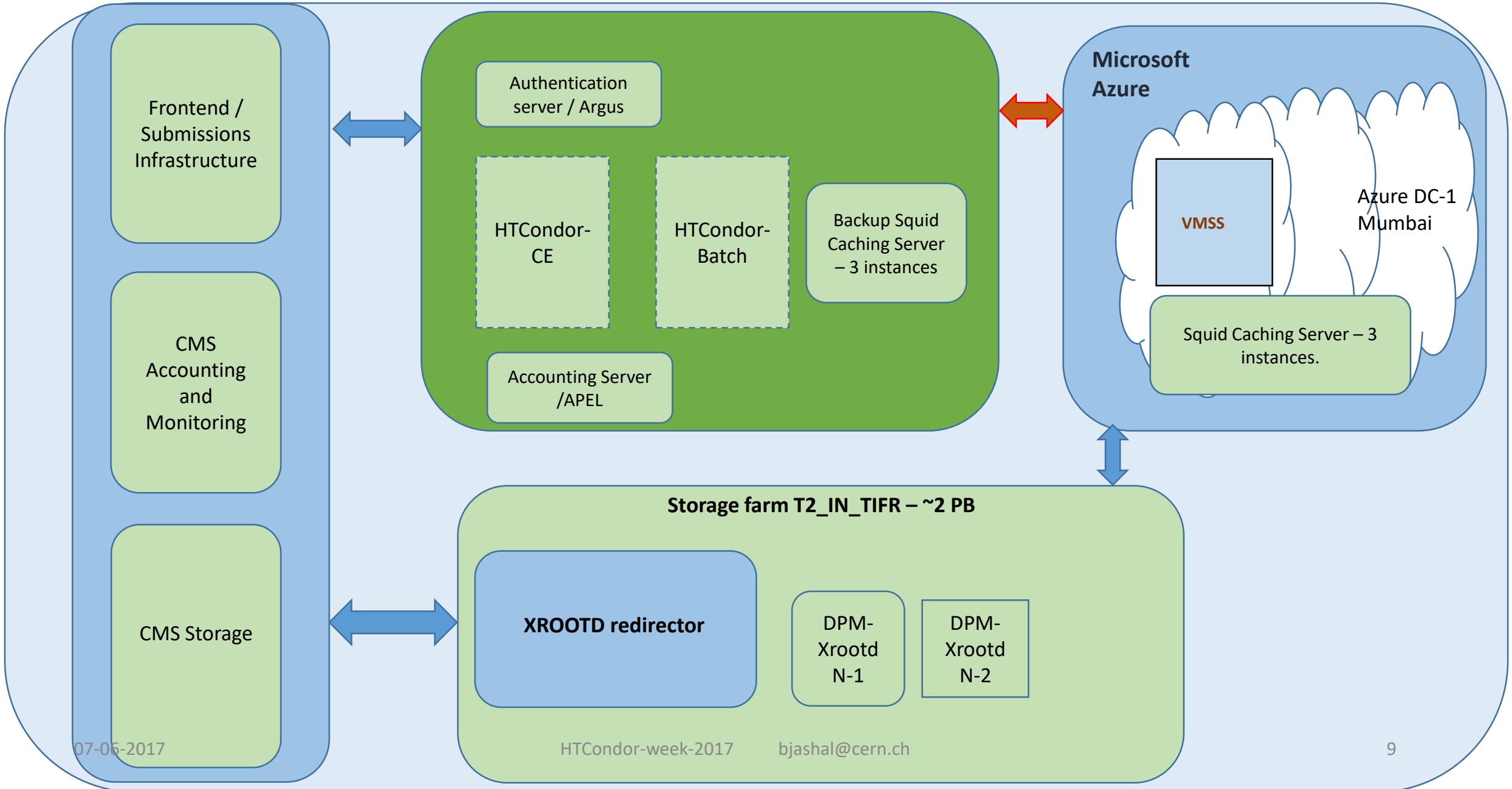
- No pre-placement of data
- Diskless site, no cloud storage used
- Stage-in and stage-out directly via TIFR xrootd redirector to any CMS site.
- No special connectivity, communication over internet.



CMS GlideInWMS entry

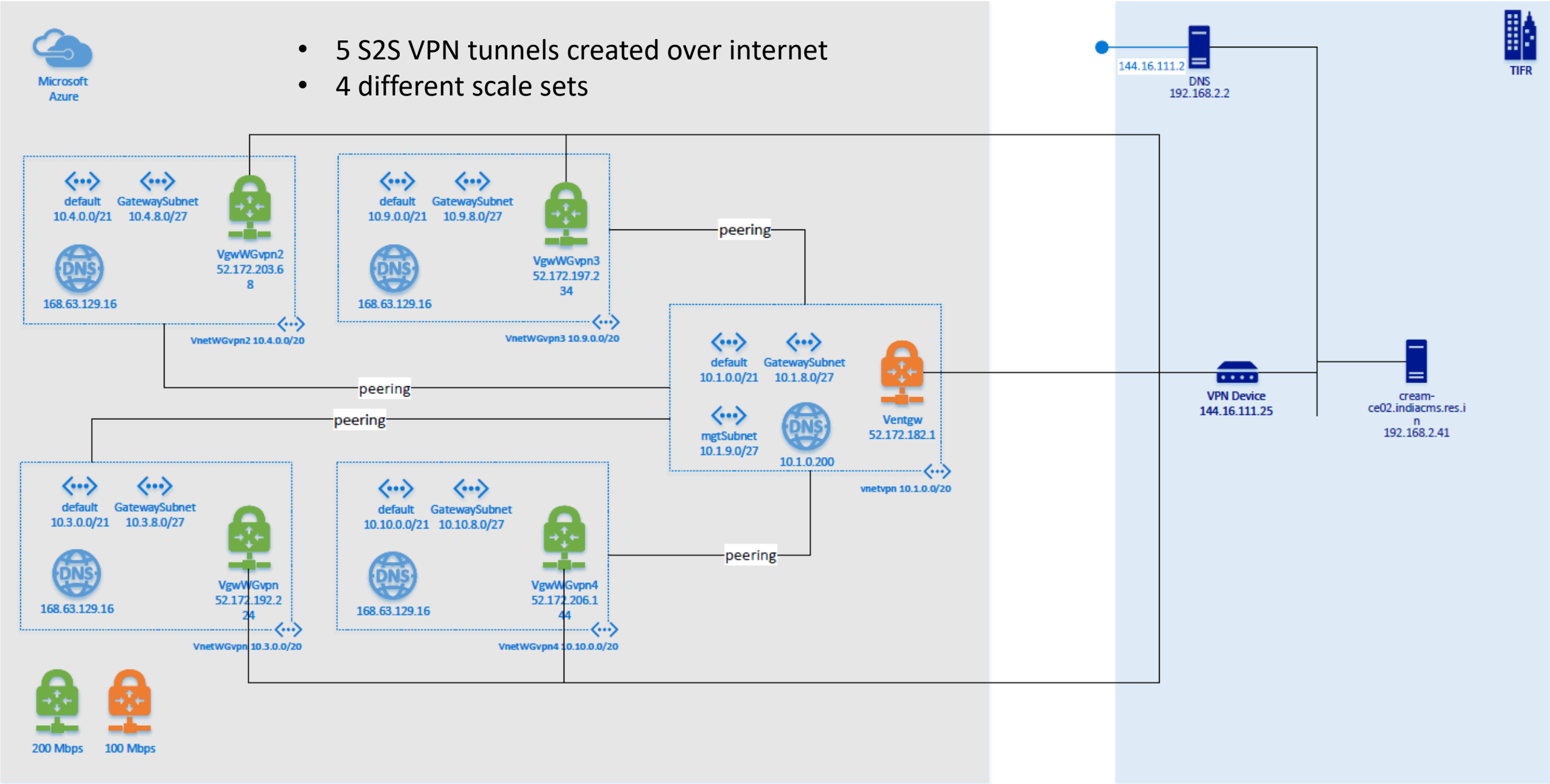
Reference diagram : USCMS

T3_IN_TIFRCloud



Azure side.

- 5 S2S VPN tunnels created over internet
- 4 different scale sets



HTCondor ⇔ Azure interfacing

- Phase One (Finished)
  Ad hoc scripts

- Phase Two (Production ready) –
  GAHP released on website - Azure GAHP

- Phase Three (In progress)
  Condor Annex
  Azure Batch

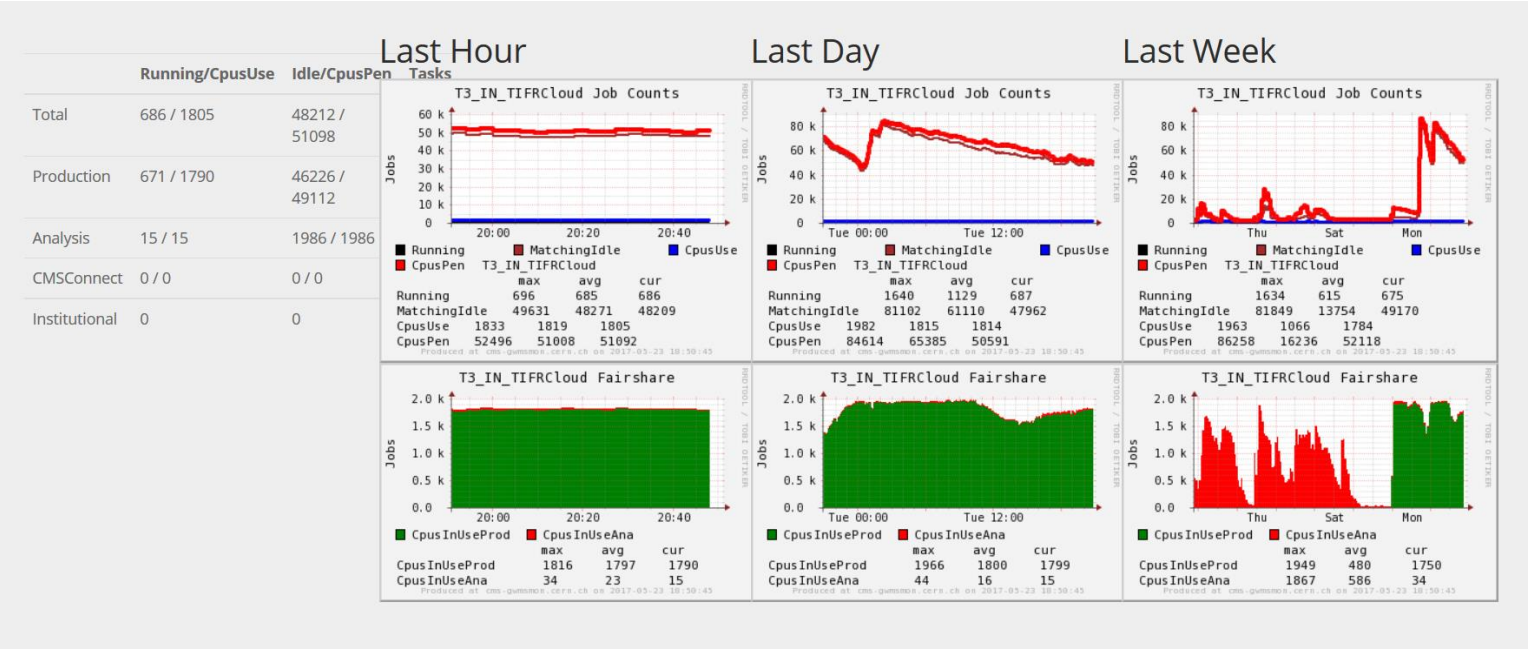
Ad hoc scripts for

1 – Auto-spinning of VMs based on load queue size

0 to 2K cores under 10 minutes.

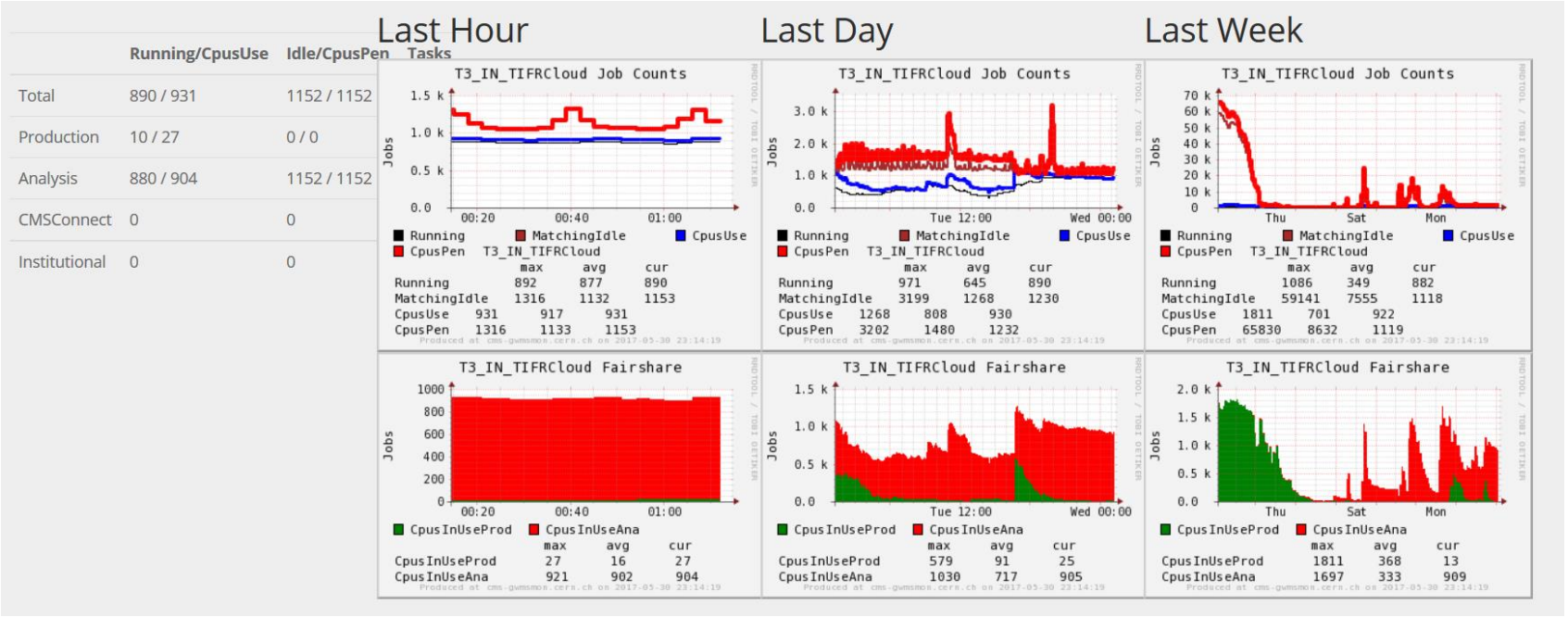
2 – Auto de-allocation of scale-setup based on queue and idle slots

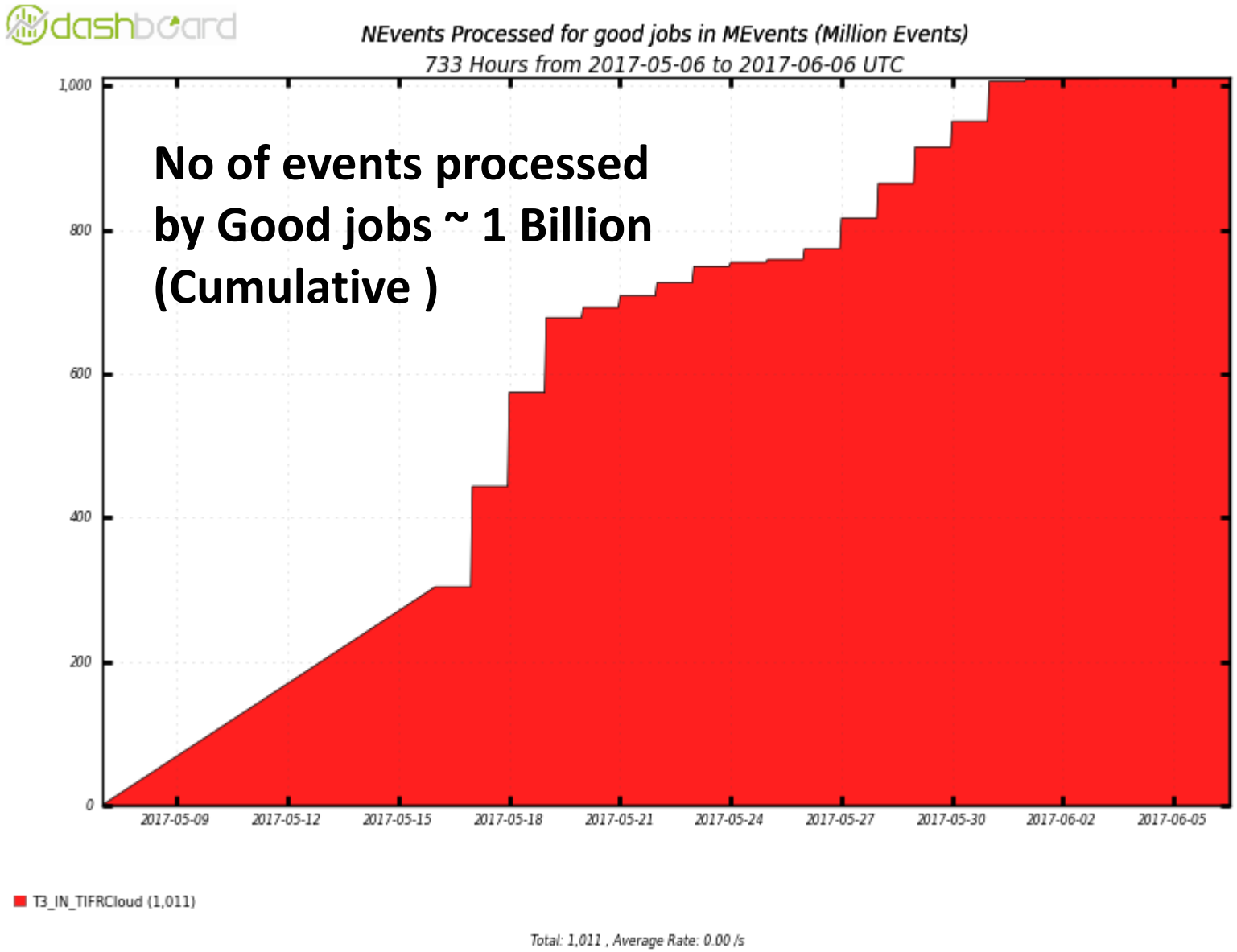
```
#To get top VM scale set Name

        vmss=`condor_status -wide | awk '{print $1,$4,$5,$8}'| awk -F'@' '{print
$2}'| awk '{if ($2=="Unclaimed" && $3 == "Idle" && $4 >= "0+00:40:00
") {print $0}}' | sort |  awk '{print $1}' | sort | uniq -c | awk '{if ($1 == "16")
{print $0}}'| awk 'NR==1{print substr($2,1); }' | cut -c 1-9`

#To get top instance ID from Condor status

        instanceid=`condor_status -wide | awk '{print $1,$4,$5,$8}'| awk -F'@'
'{print $2}'| awk '{if ($2=="Unclaimed" && $3 == "Idle" && $4 >= "0+00
:40:00") {print $0}}' | sort |  awk '{print $1}' | sort | uniq -c | awk '{if ($1 ==
"16") {print $0}}'| awk 'NR==1{print substr($2,10); }' | cut -c 1-6
`
        echo "$instanceid"

# create a character array
       arr=()
       i=0
       while [ "$i" -lt "${#instanceid}" ]; do
        arr+=("${instanceid:$i:1}")
        i=$((i+1))
       done
       echo "this is the array value"
```



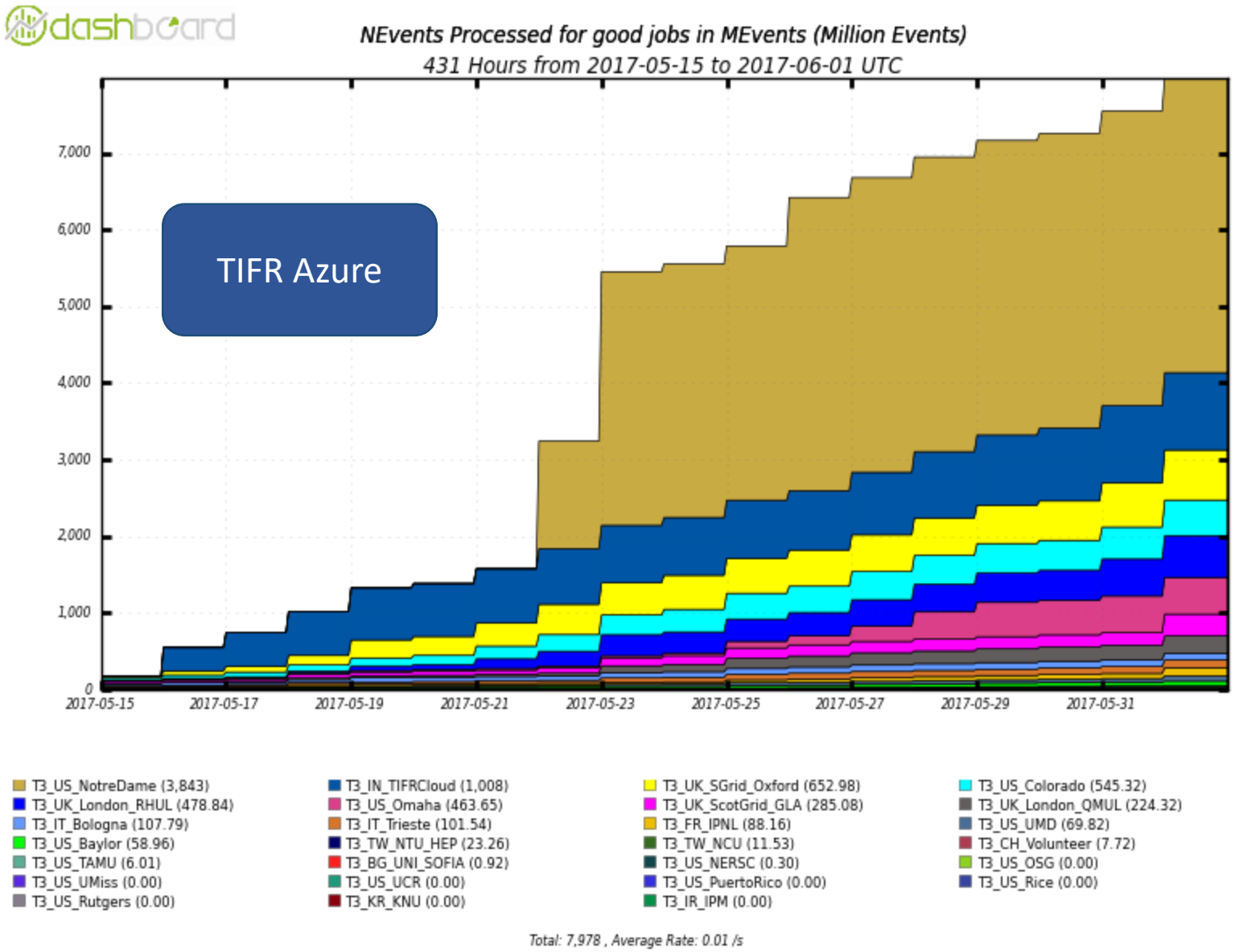
Moderate I/O production jobs reading / writing to T2_IN_TIFR and CERN

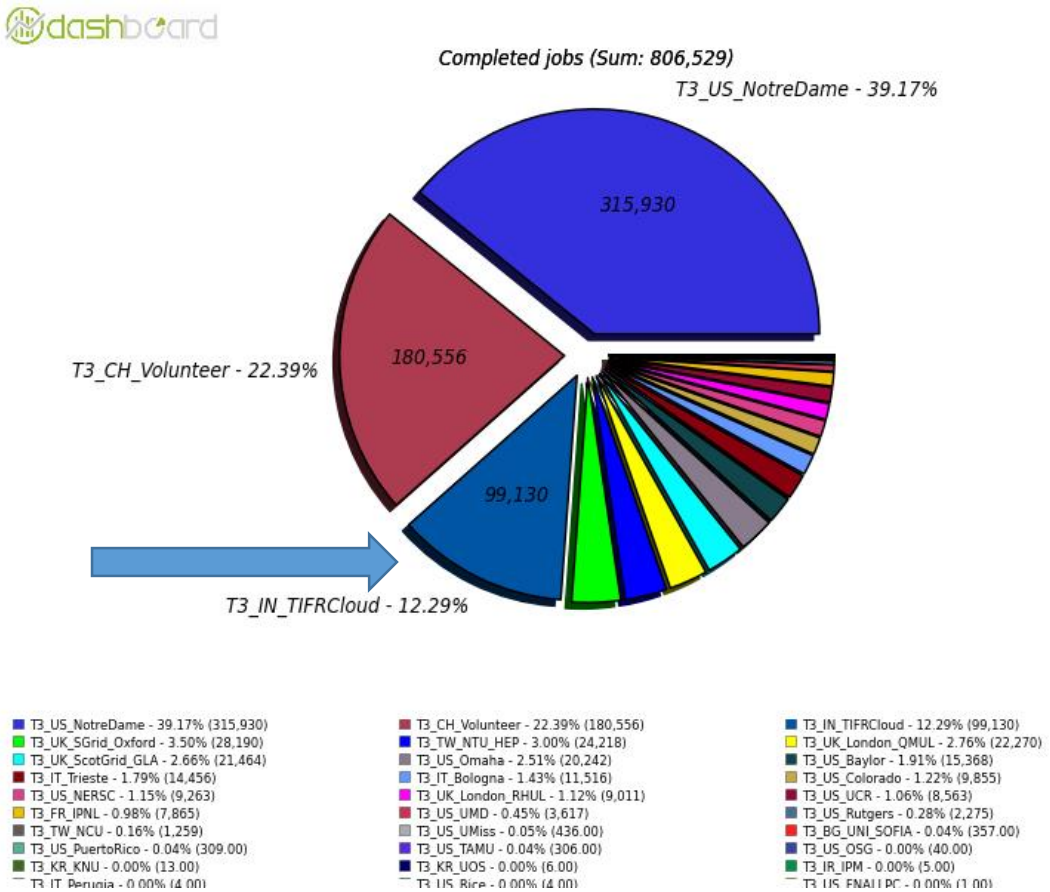
High I/O analysis jobs reading and writing from CMS xrootd sites from all over.

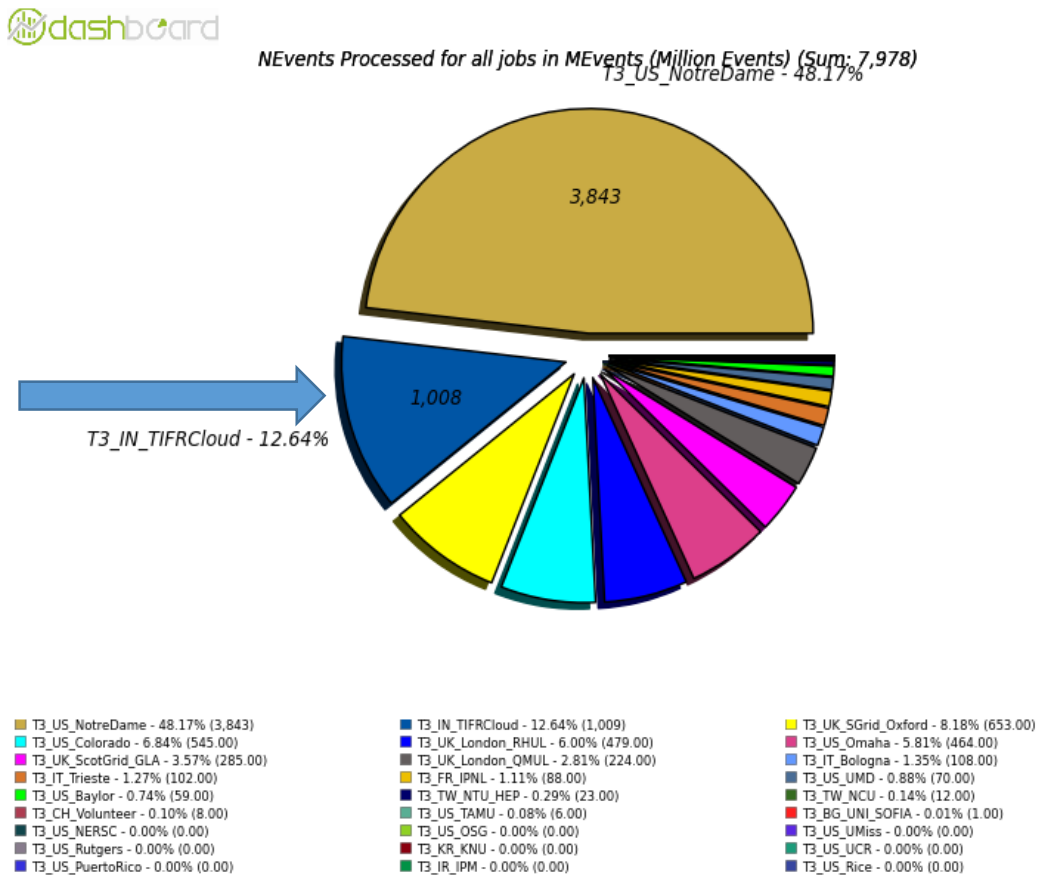




No of events processed by Good jobs ~ 1 Billion (Cumulative)


dashboard
NEvents Processed for good jobs in MEvents (Million Events)
431 Hours from 2017-05-15 to 2017-06-01 UTC
TIFR Azure
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
2017-05-15
2017-05-17
2017-05-19
2017-05-21
2017-05-23
2017-05-25
2017-05-27
2017-05-29
2017-05-31
T3_US_NotreDame (3,843)
T3_IN_TIFRCloud (1,008)
T3_UK_SGrid_Oxford (652.98)
T3_US_Colorado (545.32)
T3_UK_London_RHUL (478.84)
T3_US_Omaha (463.65)
T3_UK_ScotGrid_GLA (285.08)
T3_UK_London_QMUL (224.32)
T3_IT_Bologna (107.79)
T3_IT_Trieste (101.54)
T3_FR_IPNL (88.16)
T3_US_UMD (69.82)
T3_US_Baylor (58.96)
T3_TW_NTU_HEP (23.26)
T3_TW_NCU (11.53)
T3_CH_Volunteer (7.72)
T3_US_TAMU (6.01)
T3_BG_UNI_SOFIA (0.92)
T3_US_NERSC (0.30)
T3_US_OSG (0.00)
T3_US_UMiss (0.00)
T3_US_UCR (0.00)
T3_US_PuertoRico (0.00)
T3_US_Rice (0.00)
T3_US_Rutgers (0.00)
T3_KR_KNU (0.00)
T3_IR_IPM (0.00)
Total: 7,978 , Average Rate: 0.01 /s


dashboard
Completed jobs (Sum: 806,529)
T3_US_NotreDame - 39.17%
315,930
180,556
T3_CH_Volunteer - 22.39%
99,130
T3_IN_TIFRCloud - 12.29%
T3_US_NotreDame - 39.17% (315,930)
T3_CH_Volunteer - 22.39% (180,556)
T3_IN_TIFRCloud - 12.29% (99,130)
T3_UK_SGrid_Oxford - 3.50% (28,190)
T3_TW_NTU_HEP - 3.00% (24,218)
T3_UK_London_QMUL - 2.76% (22,270)
T3_UK_ScotGrid_GLA - 2.66% (21,464)
T3_US_Omaha - 2.51% (20,242)
T3_US_Baylor - 1.91% (15,368)
T3_IT_Trieste - 1.79% (14,456)
T3_IT_Bologna - 1.43% (11,516)
T3_US_Colorado - 1.22% (9,855)
T3_US_NERSC - 1.15% (9,263)
T3_UK_London_RHUL - 1.12% (9,011)
T3_US_UCR - 1.06% (8,563)
T3_FR_IPNL - 0.98% (7,865)
T3_US_UMD - 0.45% (3,617)
T3_US_Rutgers - 0.28% (2,275)
T3_TW_NCU - 0.16% (1,259)
T3_US_UMiss - 0.05% (436.00)
T3_BG_UNI_SOFIA - 0.04% (357.00)
T3_US_PuertoRico - 0.04% (309.00)
T3_US_TAMU - 0.04% (306.00)
T3_US_OSG - 0.00% (40.00)
T3_KR_KNU - 0.00% (13.00)
T3_KR_UOS - 0.00% (6.00)
T3_IR_IPM - 0.00% (5.00)
T3_IT_Perugia - 0.00% (4.00)
T3_US_Rice - 0.00% (4.00)
T3_US_ENAU_PC - 0.00% (1.00)


dashboard
NEvents Processed for all jobs in MEvents (Million Events) (Sum: 7,978)
T3_US_NotreDame - 48.17%
3,843
1,008
T3_IN_TIFRCloud - 12.64%
T3_US_NotreDame - 48.17% (3,843)
T3_IN_TIFRCloud - 12.64% (1,009)
T3_UK_SGrid_Oxford - 8.18% (653.00)
T3_US_Colorado - 6.84% (545.00)
T3_UK_London_RHUL - 6.00% (479.00)
T3_US_Omaha - 5.81% (464.00)
T3_UK_ScotGrid_GLA - 3.57% (285.00)
T3_UK_London_QMUL - 2.81% (224.00)
T3_IT_Bologna - 1.35% (108.00)
T3_IT_Trieste - 1.27% (102.00)
T3_FR_IPNL - 1.11% (88.00)
T3_US_UMD - 0.88% (70.00)
T3_US_Baylor - 0.74% (59.00)
T3_TW_NTU_HEP - 0.29% (23.00)
T3_TW_NCU - 0.14% (12.00)
T3_CH_Volunteer - 0.10% (8.00)
T3_US_TAMU - 0.08% (6.00)
T3_BG_UNI_SOFIA - 0.01% (1.00)
T3_US_NERSC - 0.00% (0.00)
T3_US_OSG - 0.00% (0.00)
T3_US_UMiss - 0.00% (0.00)
T3_US_Rutgers - 0.00% (0.00)
T3_KR_KNU - 0.00% (0.00)
T3_US_UCR - 0.00% (0.00)
T3_US_PuertoRico - 0.00% (0.00)
T3_IR_IPM - 0.00% (0.00)
T3_US_Rice - 0.00% (0.00)

Network usage:

Estimated network B/W requirement

- .5 Mbps per core ➔ for 2K cores = 1Gbps.
- Before starting the run, several tests of upto 2Gbps transfers from TIFR to Azure via shared internet link, B/W not guaranteed

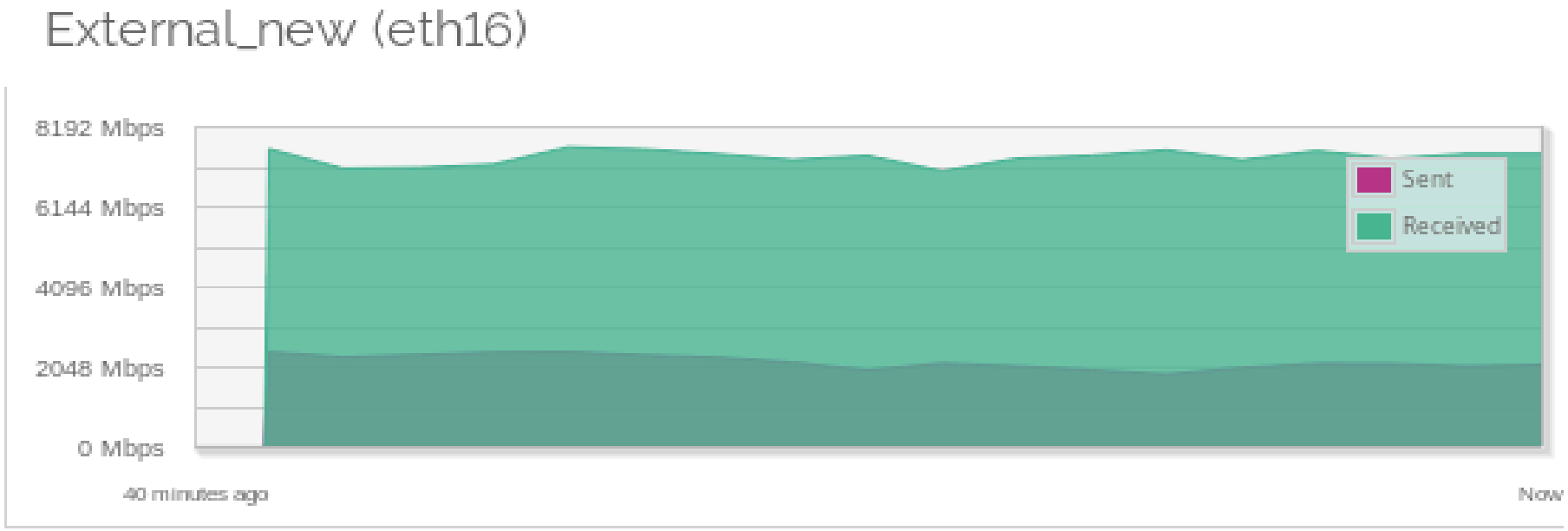
Actual utilization during the Run => Read - Average ~ 2.5Gbps , Peak ~ 4Gbps

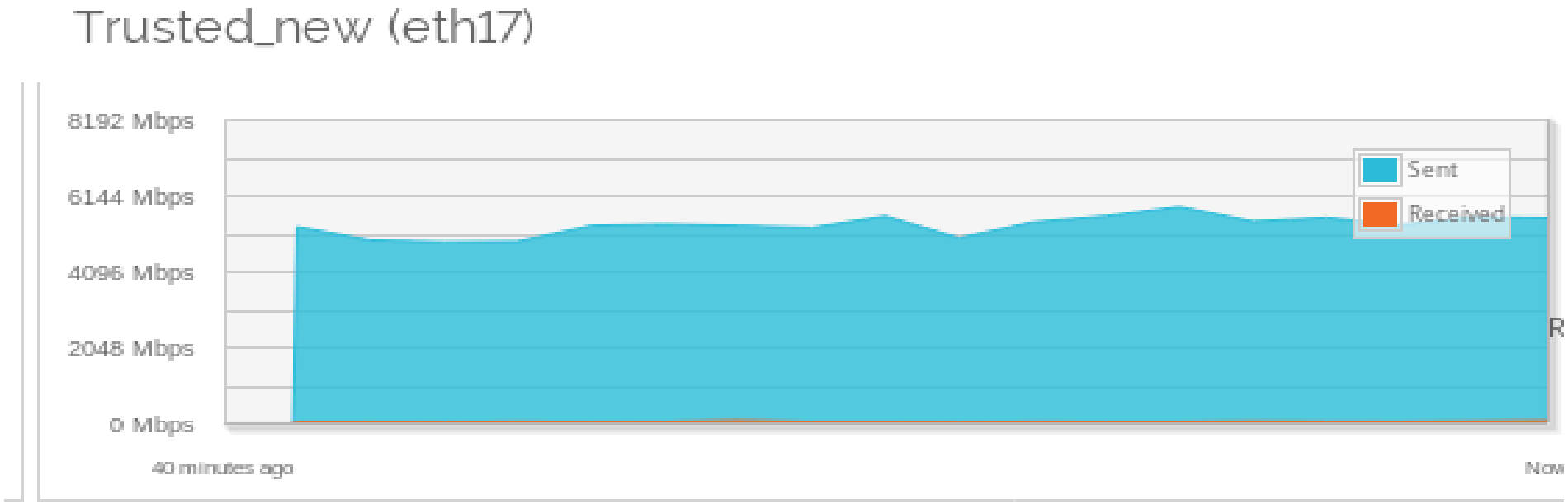
Write – Average ~ 1.7 Gbps, Peak ~ 2.6 Gbps

Storage usage:

Reached limit with 12K network threads from jobs running all over along with cloud cores and T2 cores.

Network usage:





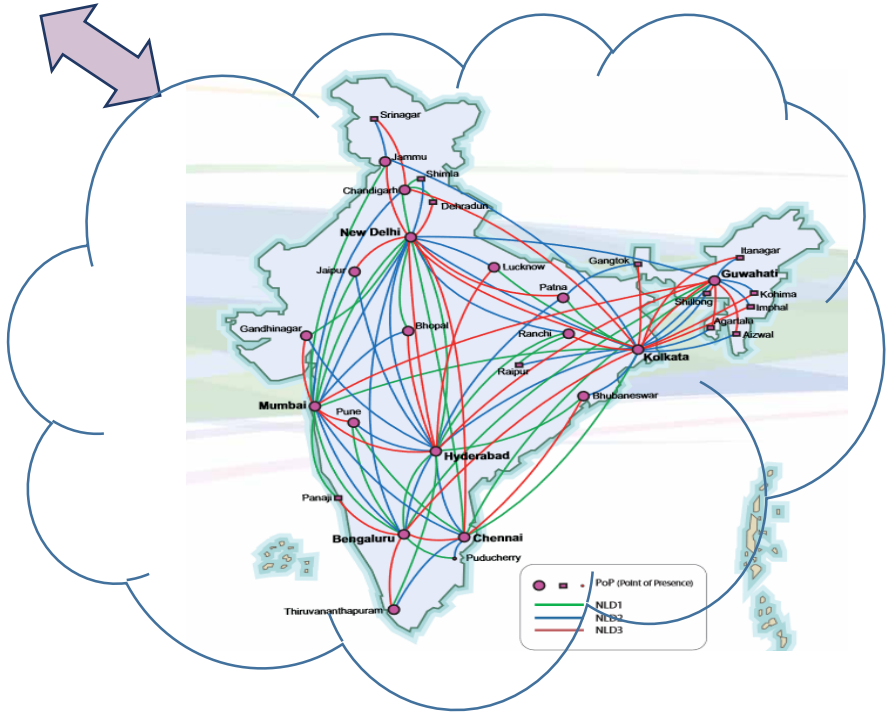
- HTCondor pool of all India-CMS collaborating institutes.
- Bringing resources time to time from Indian institutes and various cloud platforms

Collaborating Indian Institutes

- **TIFR, Mumbai** WLCG Site
- **VECC, Kolkata** WLCG Site (Alice)
- BARC, Mumbai
- Delhi University, New Delhi
- SINP, Kolkata
- Punjab University, Chandigarh
- RRCAT, Indore
- IOP, Bhubneshwar
- IISER, Pune
- Rajasthan university
- **IIT, Chennai**
- NISER, Bhubneshwar
- IISC Banglore

CMS – LHCONE with 10G => 40G

L3VRF of all collaborating institutes over NKN



NATIONAL SUPERCOMPUTING MISSION

INFRASTRUCTURE | APPLICATIONS | R&D | HRD

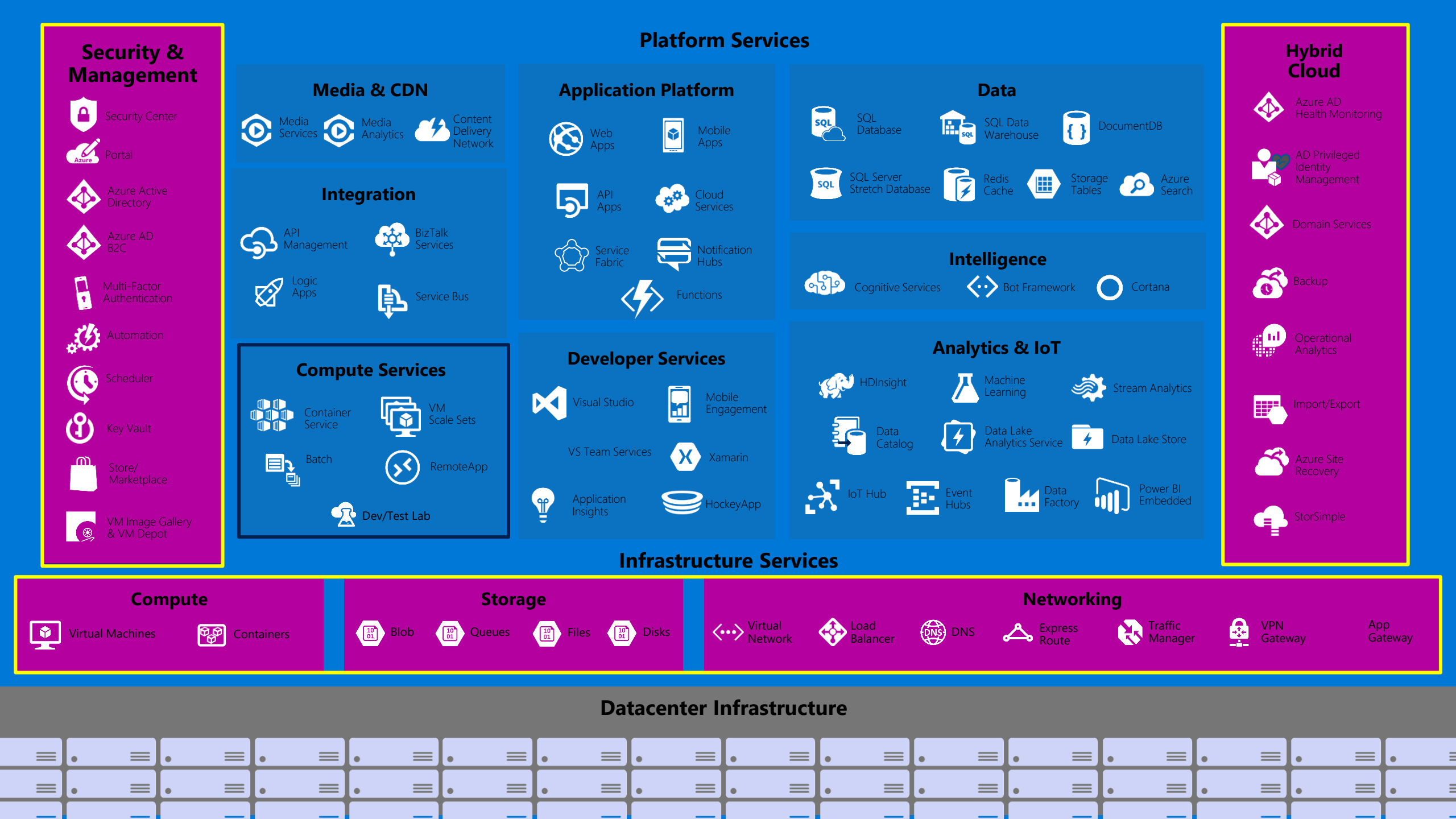
- Budget overlay of ~700 Million USD
- To be invested in Infrastructure, Applications, R&D and HRD
- 70 HPC / HTC computing facilities.
- Heterogeneous workloads from different sciences and users.

Microsoft Azure

Problem: Submit CERN jobs to the Microsoft Cloud, Azure

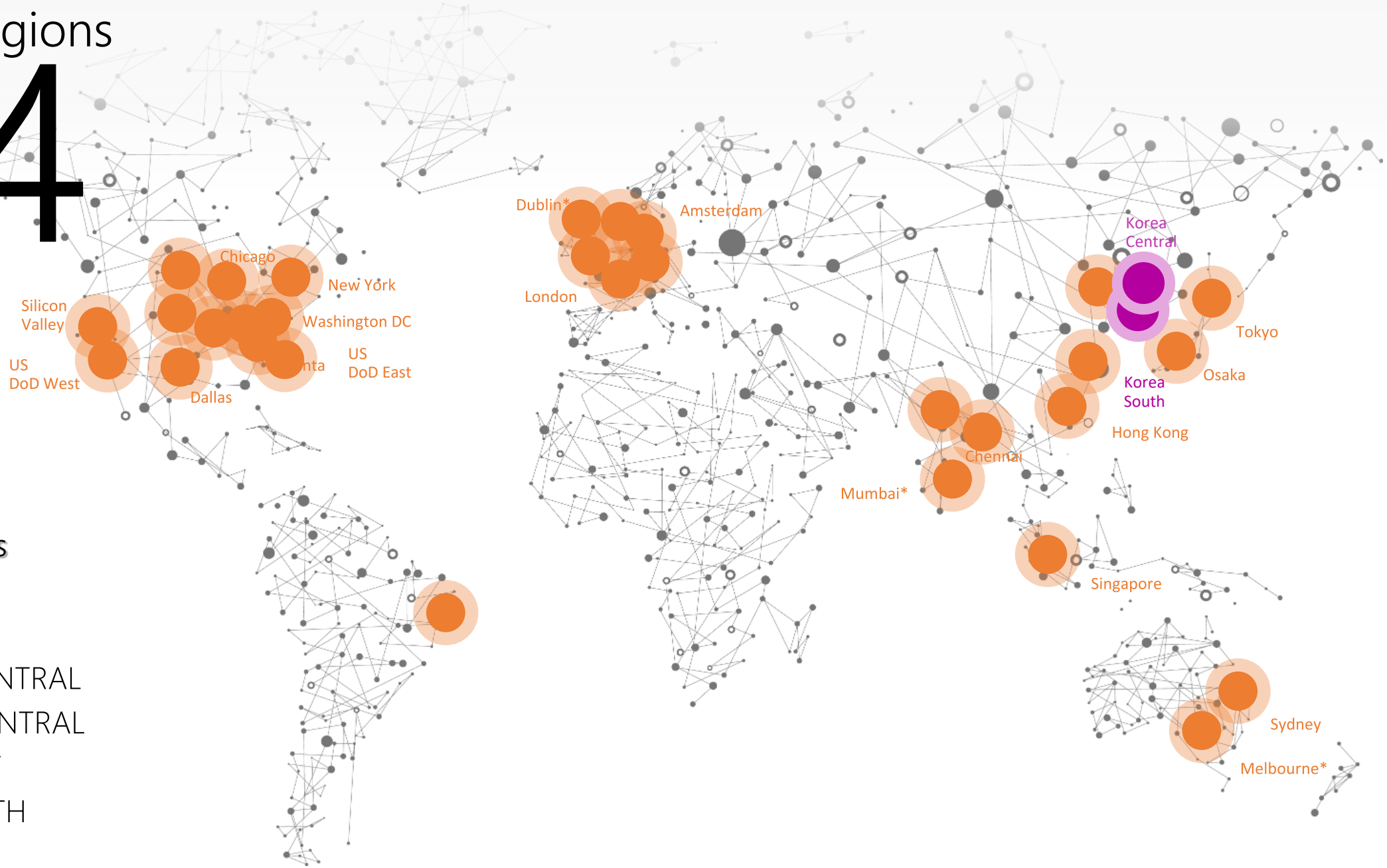
Goal: Enable HTCondor to schedule on Azure, seamlessly through Grid Universe

Technology: Azure H-series, embarrassingly parallel, ARM, API


Security & Management
Security Center
Portal
Azure Active Directory
Azure AD B2C
Multi-Factor Authentication
Automation
Scheduler
Key Vault
Store/ Marketplace
VM Image Gallery & VM Depot
Platform Services
Media & CDN
Media Services
Media Analytics
Content Delivery Network
Integration
API Management
BizTalk Services
Logic Apps
Service Bus
Compute Services
Container Service
VM Scale Sets
Batch
RemoteApp
Dev/Test Lab
Application Platform
Web Apps
Mobile Apps
API Apps
Cloud Services
Service Fabric
Notification Hubs
Functions
Developer Services
Visual Studio
Mobile Engagement
VS Team Services
Xamarin
Application Insights
HockeyApp
Data
SQL Database
SQL Data Warehouse
DocumentDB
SQL Server Stretch Database
Redis Cache
Storage Tables
Azure Search
Intelligence
Cognitive Services
Bot Framework
Cortana
Analytics & IoT
HDInsight
Machine Learning
Stream Analytics
Data Catalog
Data Lake Analytics Service
Data Lake Store
IoT Hub
Event Hubs
Data Factory
Power BI Embedded
Hybrid Cloud
Azure AD Health Monitoring
AD Privileged Identity Management
Domain Services
Backup
Operational Analytics
Import/Export
Azure Site Recovery
StorSimple
Infrastructure Services
Compute
Virtual Machines
Containers
Storage
Blob
Queues
Files
Disks
Networking
Virtual Network
Load Balancer
DNS
Express Route
Traffic Manager
VPN Gateway
App Gateway
Datacenter Infrastructure

Azure regions

34

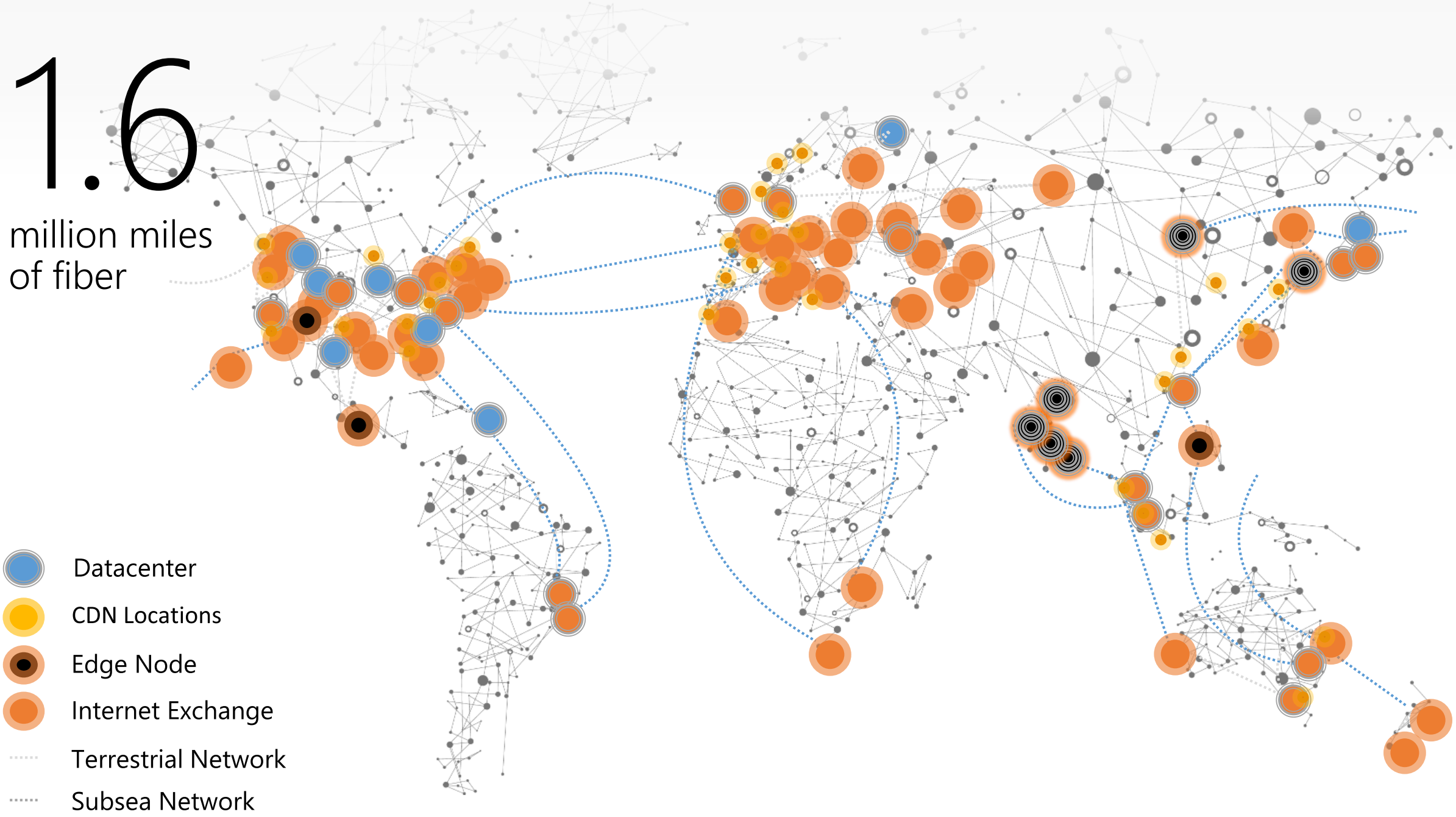


HPC/GPU regions

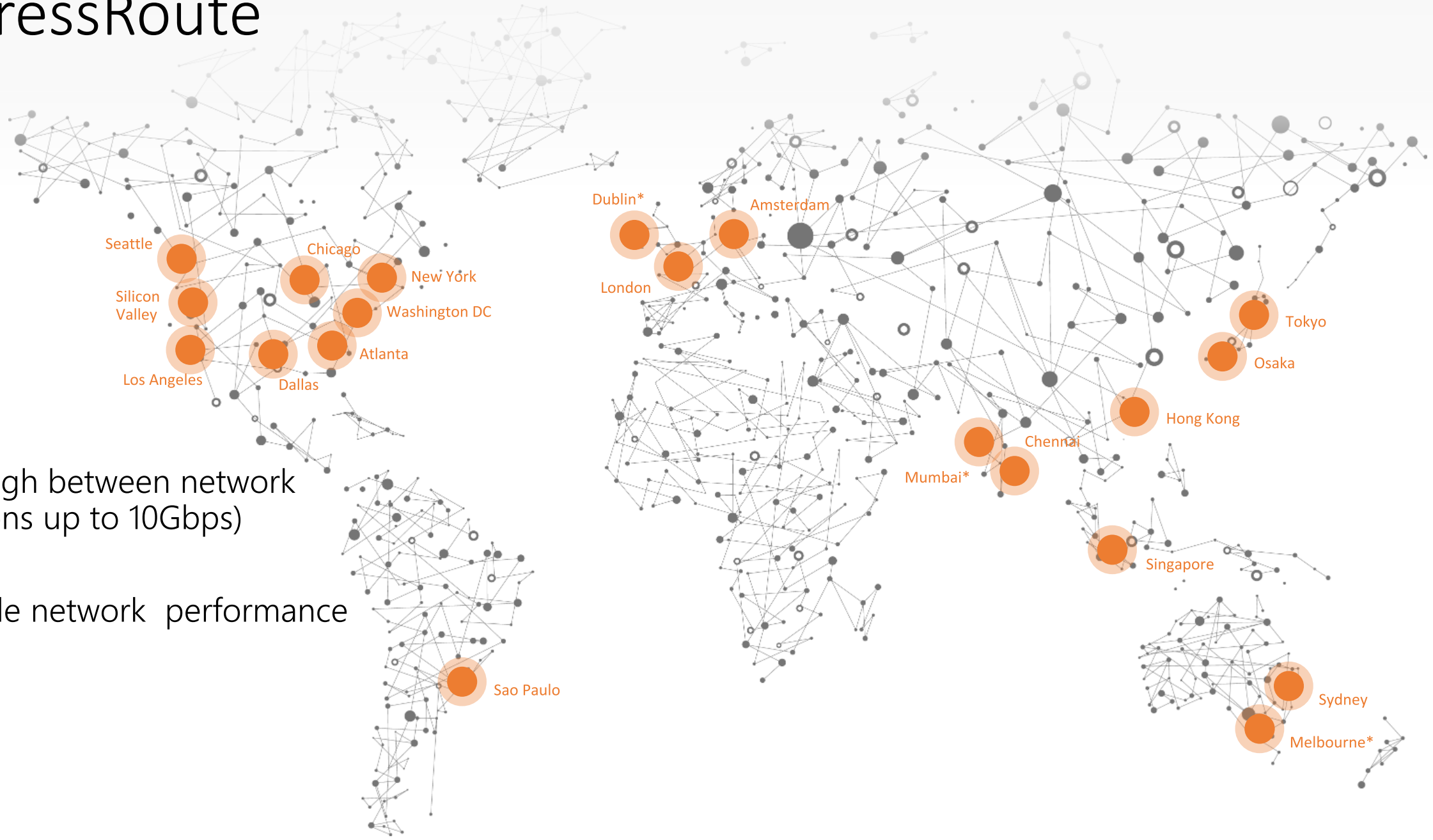
- US EAST
- US WEST
- US SOUTH CENTRAL
- US NORTH CENTRAL
- EUROPE WEST
- EUROPE NORTH
- JAPAN EAST

1.6

million miles
of fiber

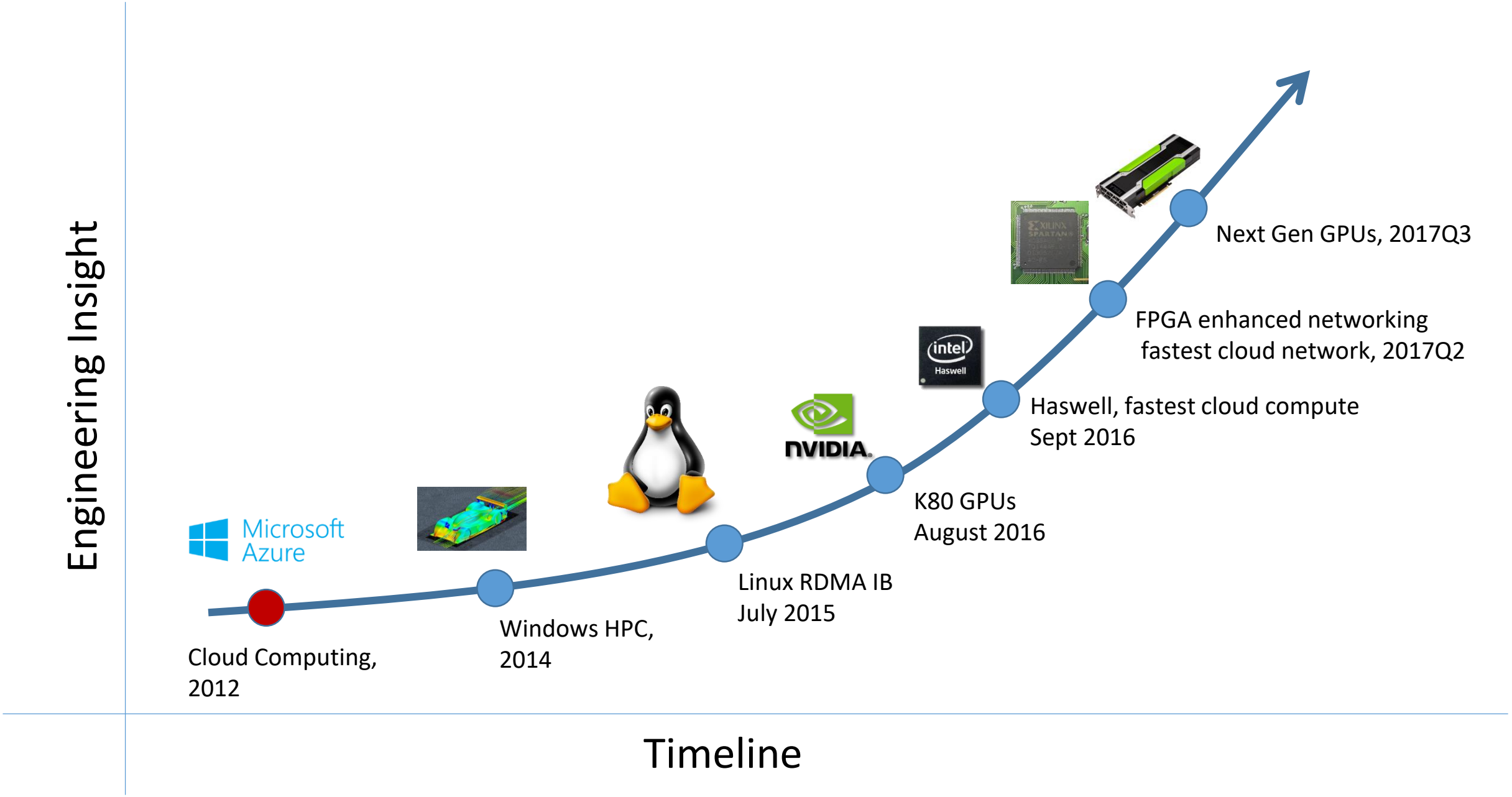


ExpressRoute



Private, high between network connections up to 10Gbps)

Predictable network performance


Engineering Insight
Microsoft Azure
Cloud Computing, 2012
Windows HPC, 2014
Linux RDMA IB July 2015
NVIDIA
K80 GPUs August 2016
intel Haswell
Haswell, fastest cloud compute Sept 2016
FPGA enhanced networking fastest cloud network, 2017Q2
Next Gen GPUs, 2017Q3
Timeline

Azure is an open cloud

DevOps

Clients

Management

Schedulers

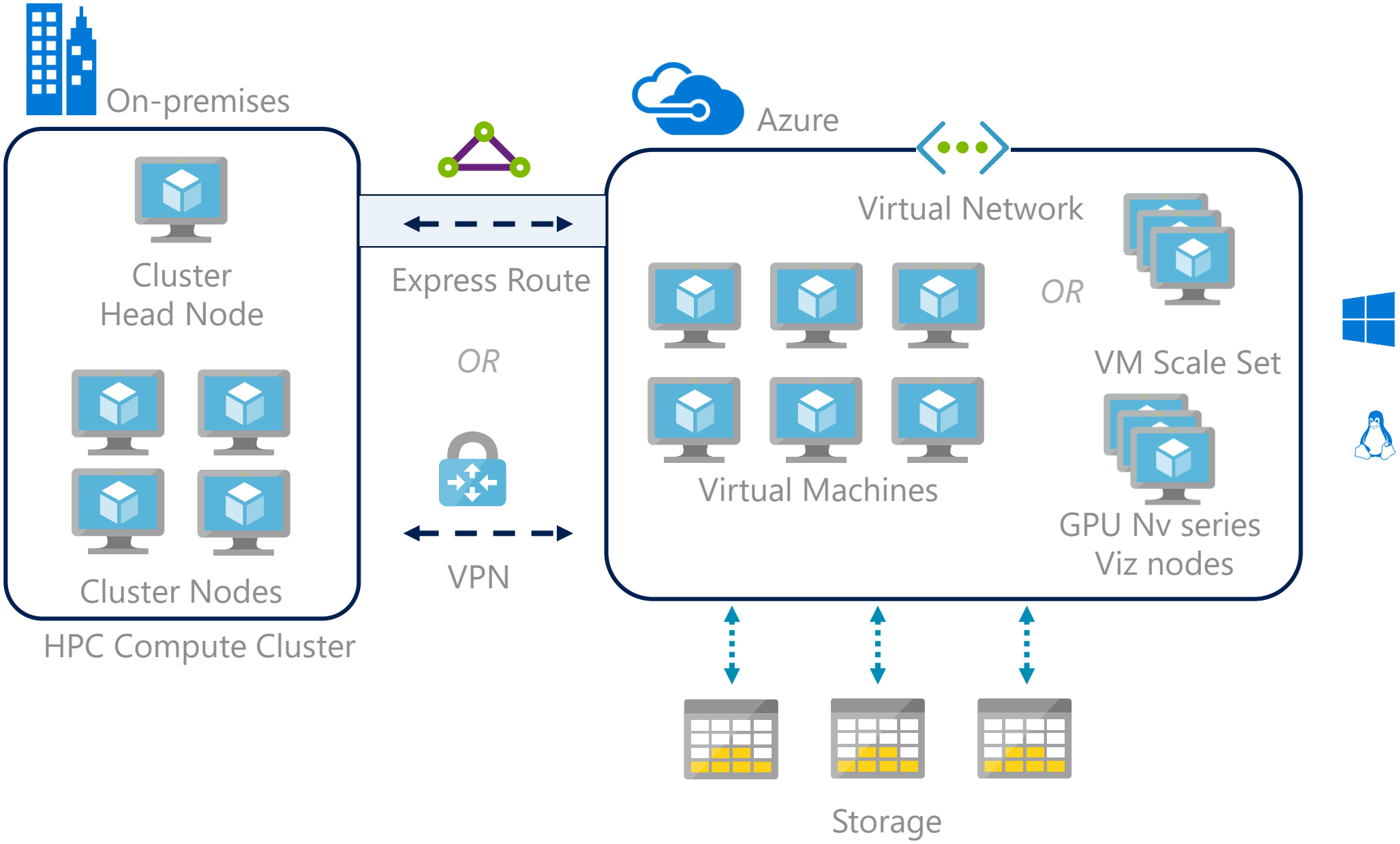
PaaS and DevOps

App frameworks and tools

Databases and middleware

Infrastructure

Architecture for Hybrid Cloud



Azure Resource Management (ARM) Templates

```
{
    "$schema": "http://schema.management.azure.com/schemas/2015-01-01/deplo
    "contentVersion": "",
    "parameters": {
        "<parameter-name>" : {
            "type" : "<type-of-parameter-value>",
            "defaultValue": "<default-value-of-parameter>",
            "allowedValues": [ "<array-of-allowed-values>" ],
            "minValue": <minimum-value-for-int>,
            "maxValue": <maximum-value-for-int>,
            "minLength": <minimum-length-for-string-or-array>,
            "maxLength": <maximum-length-for-string-or-array-parameters>,
            "metadata": {
                "description": "<description-of-the parameter>"
            }
        }
    },
    "variables": {
        "<variable-name>": "<variable-value>",
        "<variable-name>": {
            <variable-complex-type-value>
        }
    },
    "resources": [
      {
          "apiVersion": "<api-version-of-resource>",
          "type": "<resource-provider-namespace/resource-type-name>",
          "name": "<name-of-the-resource>",
          "location": "<location-of-resource>",
          "tags": "<name-value-pairs-for-resource-tagging>",
          "comments": "<your-reference-notes>",
          "dependsOn": [
              "<array-of-related-resource-names>"
          ],
          "properties": "<settings-for-the-resource>",
```

https://github.com/xpillons/azure-hpc/

TIFR's use of HTCondor in Azure

- Allocate HTCondor pool in Azure
- Phase One (finished)
  - Ad hoc scripts
- Phase Two (in progress)
  - Grid universe
- Phase Three (future work)
  - Condor Annex
  - Azure Batch?

Azure Grid Universe

- New resource type in grid universe
  - Similar to existing support for EC2, GCE
- Each VM is a job
- Works well with glide-in factories
  - GlideinWMS
- Working prototype
  - Will be in HTCondor 8.7.x
  - Not highly scalable, hope to improve

Sample Submit File

```
universe = grid
grid_resource = azure \
   837155e9-a033-488c-8742-645cefa38c89
executable = jfrey-test
azure_auth_file = azure-jfrey.creds
azure_location = centralindia
azure_size = Standard_DS1_V2
azure_image = linux-ubuntu-latest
azure_custom_data = install-condor.dat
azure_admin_username = jfrey
azure_admin_key = azure_rsa.pub
queue
```

Future Work

- Condor Annex
  - Better for high scalability
  - Use VM Scale Sets for efficiency
  - Easier management of many identical VMs
- Azure Batch
  - Submit “regular” jobs to Azure’s job scheduler

Microsoft

© Microsoft Corporation. All rights reserved.