Azure for Research

Mark McManus markmcm@microsoft.com



University of Cambridge and Microsoft Research

Deep engagements across computer science, engineering, health and life sciences

Funded **70 PhD students at the University of Cambridge (since 2004)**and hosted **45 PhD interns (since 2016).**

Many MSR researchers are alumni of the University, and Fellows of Colleges.

Health Data Research UK Team of the Year 2020 – with Cambridge University
and Microsoft - <u>HDR UK's Team of the</u>
Year award 2020 - HDR UK

Multi-million pound MicrosoftCambridge Machine Learning Initiative
EPSRC Prosperity Partnership kicked
off with Engineering Department - UK
government backs University of
Cambridge and Microsoft mission to
build better AI | Department of
Engineering

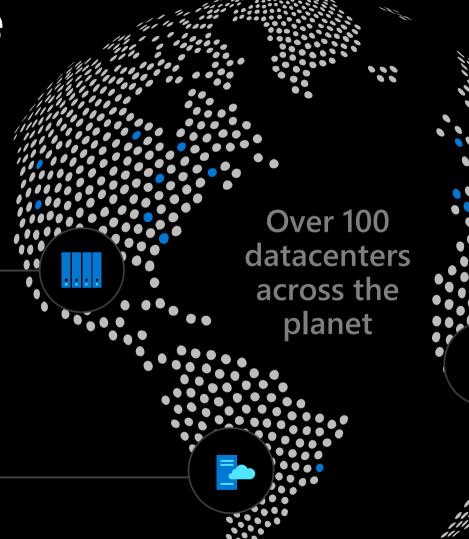
Project InnerEye AI for Cancer treatment with University of Cambrige and Addenbrookes being deployed on Azure for patients – just won £0.5M NHSX AI Award - Cambridge award winners in AI innovation | CUH



A secure foundation at global scale

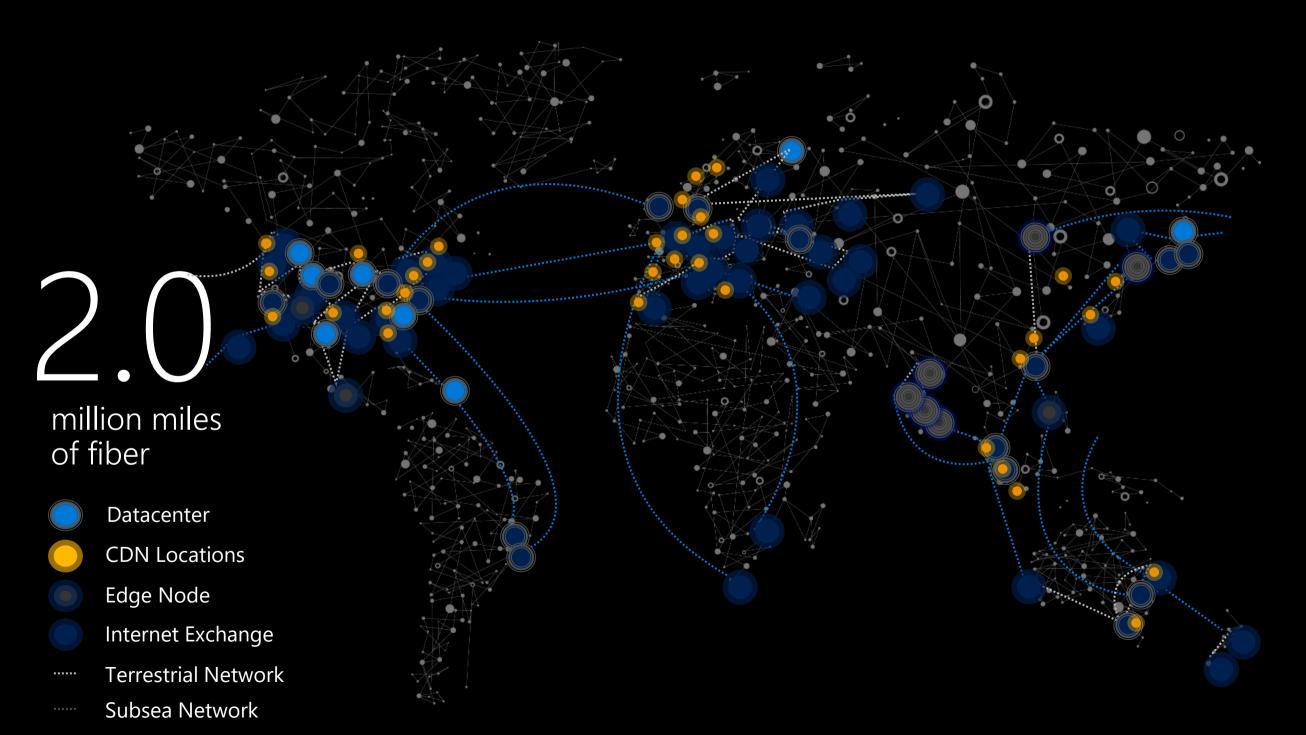
Each **physical datacenter** protected with world-class, multi-layered protection

Global cloud infrastructure with custom hardware and network protection



Secured with cutting-edge operational security

- Restricted access
- 24x7 monitoring
- Global security experts





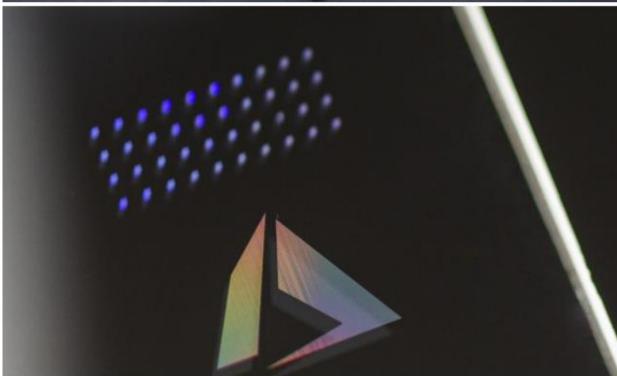
Addressing IT for Research challenges

Customer Challenges	Solution	Business Benefits	
Less than 50% of IT budget is spent through Central IT Services	Centralise IT Budget spend for Research	Reduce costs (economies of scale) Increase efficiency Less time on procurement	
Requests from Academics and Research can take 3-6 months+ to procure and deploy	On demand & Self Service environment with virtually unlimited capacity	Central IT able to service Research requirements immediately	
Research and Academic IT departmental environments typically lack governance	Centrally managed environment with Governance framework included	Central IT able to fully support Research environments Compliant Research	
Researchers and Academics spend too much time procuring and building IT environments	Central on demand services provide almost immediate access to required environment	More time for actual research Agile Research environment Faster publication	
Researchers often need short term access to expensive Research environments	Agile environment that provides required services for the time needed. No CAPEX, Only charged for actual usage	No CAPEX requirement Only charged for actual use Research Funding for IT can be held until required.	

Research Benefits

- Agile environment
- Rapid access to services
- Flexible resources adapt to project
- Access to Cognitive Services
- Cash Flow
- Enhanced security and compliance
- Industry Accreditations (ISO, NHS etc)
- Increased research capacity
- Quicker to publication
- Sustainability
- Access to new innovations, sooner







NEWCASTLE UNIVERSITY

Petascale Cloud Supercomputing for Terapixel Visualization of a Digital Twin

Nicolas S. Holliman, Member IEEE Computer Society, Manu Antony, James Charlton, Stephen Dowsland, Philip James and Mark Turner

Abstract—Background—Photo-realistic terapixel visualization is computationally intensive and to date there have been no such visualizations of urban digital twins, the few terapixel visualizations that exist have looked towards space rather than earth. Objective—our aims are: creating a scalable cloud supercomputer software architecture for visualization; a photo-realistic terapixel 3D visualization of urban IoT data supporting daily updates; a rigorous evaluation of cloud supercomputing for our application. Method—We migrated the Blender Cycles path tracer to the public cloud within a new software framework designed to scale to petaFLOP performance. Results—we demonstrate we can compute a terapixel visualization in under one hour, the system scaling at 98% efficiency to use 1024 public cloud GPU nodes delivering 14 petaFLOPS. The resulting terapixel image supports interactive browsing of the city and its data at a wide range of sensing scales. Conclusion—The GPU compute resource available in the cloud is greater than anything available on our national supercomputers providing access to globally competitive resources. The direct financial cost of access, compared to procuring and running these systems, was low. The indirect cost, in overcoming teething issues with cloud software development, should reduce significantly over time.

Index Terms—Data Visualization, Internet of Things, Scalability, Supercomputers



Azure: Trusted

Global























ISO 27001

ISO 27018

ISO 27017

ISO 22301

SOC 1 Type 2

SOC 2 Type 2

SOC 3

CSA STAR Self-Assessment

CSA STAR Certification

CSA STAR Attestation

Regional









































Argentina PDPA

Clauses

UK G-Cloud

China DJCP





Australia IRAP/CCSL

New Zealand GCIO

Japan My Number Act

ENISA IAF

Japan CS Mark Gold

Spain ĖNS

Spain DPA

India Canada MeitY Privacy Laws

Privacy Shield

Germany IT Grundschutz workbook

Industry































PCI DSS Level 1



MPAA

FACT UK

Shared Assessments FISC Japan

HIPAA/ HITECH Act

HITRUST

GxP 21 CFR Part 11 MARS-E

Section 508 VPAT

IG Toolkit UK

FERPA

GLBA

FFIEC

Us Gov







High JAB P-ATO





SRG Level 4



SRG Level 5



SP 800-171



FIPS 140-2





ITAR





IRS 1075

CJIS



Using Compliance to broaden Research capability

Azure is an open cloud

DevOps













Clients









Management





















Applications







PaaS and **DevOps**











App frameworks and tools















Databases and middleware

















Infrastructure









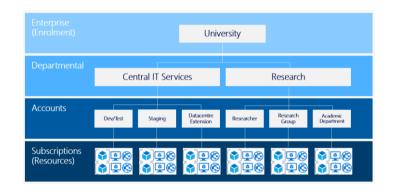






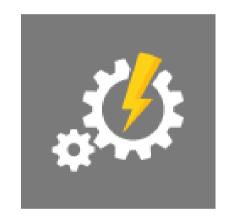


Governed service for Research



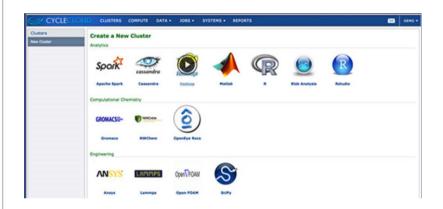


Research structure, billing, account and delegation management. Central Governance of Research services is a key aim of most Universities.



Automation for Research

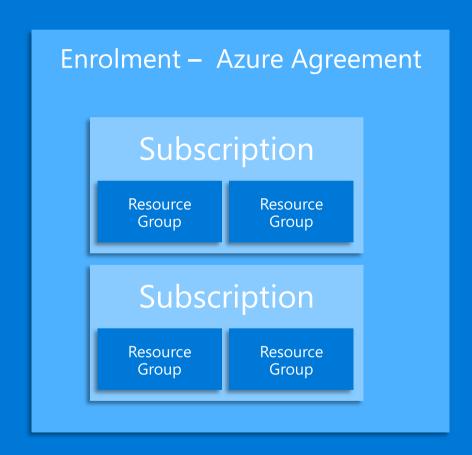
Immediate access to resources. No procurement process or pre-build saves time that can be directed to actual research

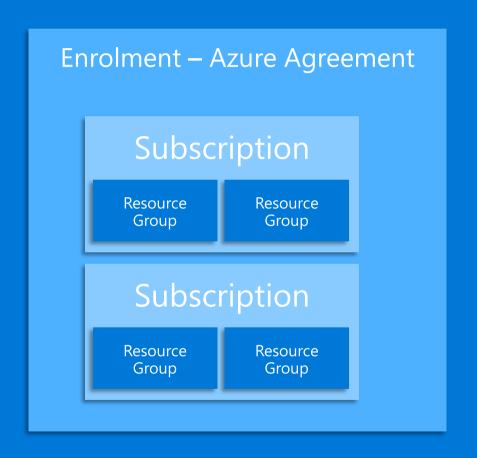


On demand Research services

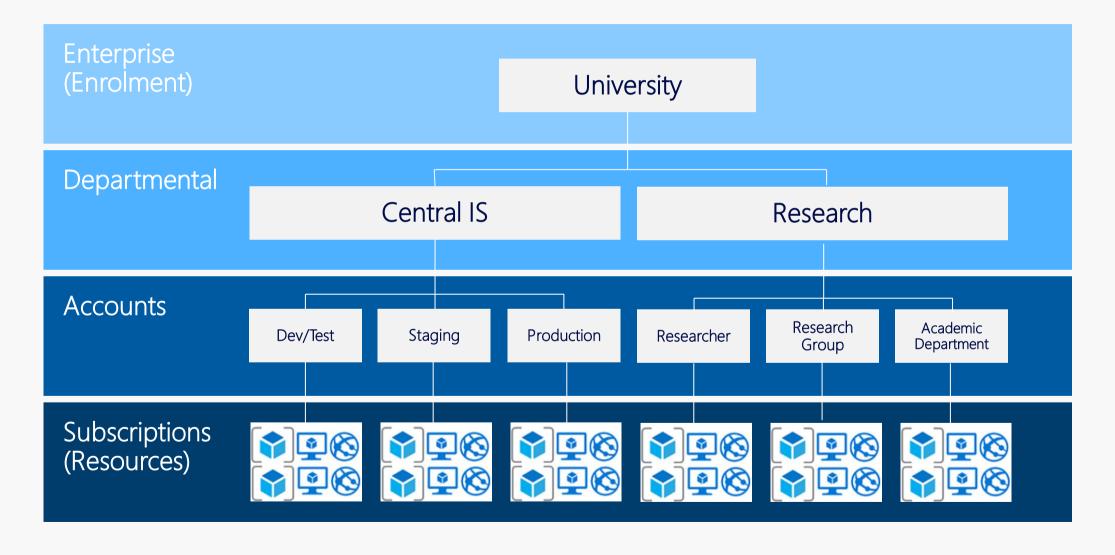
Researchers able to select their own resources without having to learn Azure or connect with Central IT

Customer Azure Environment

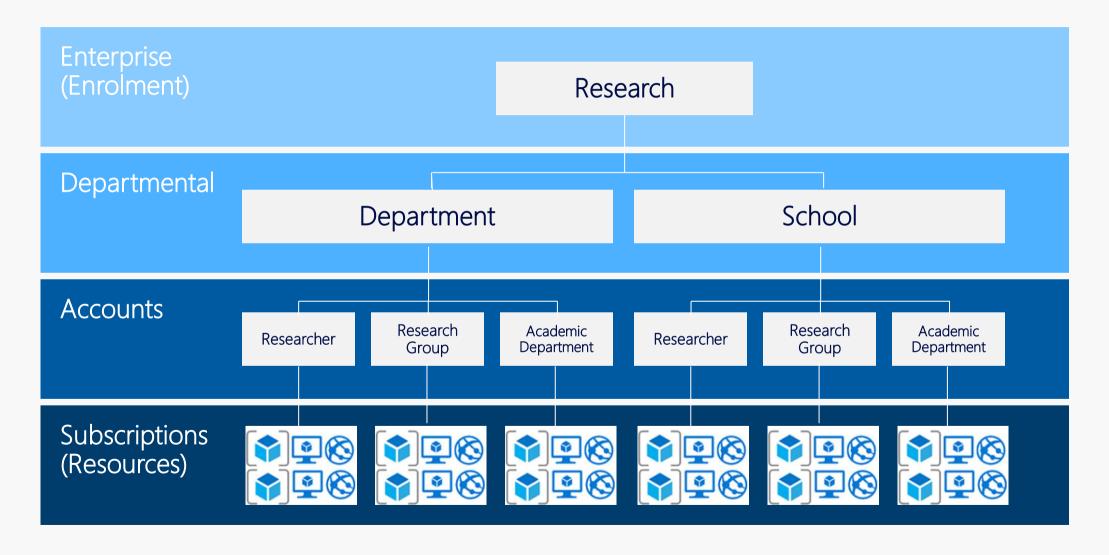


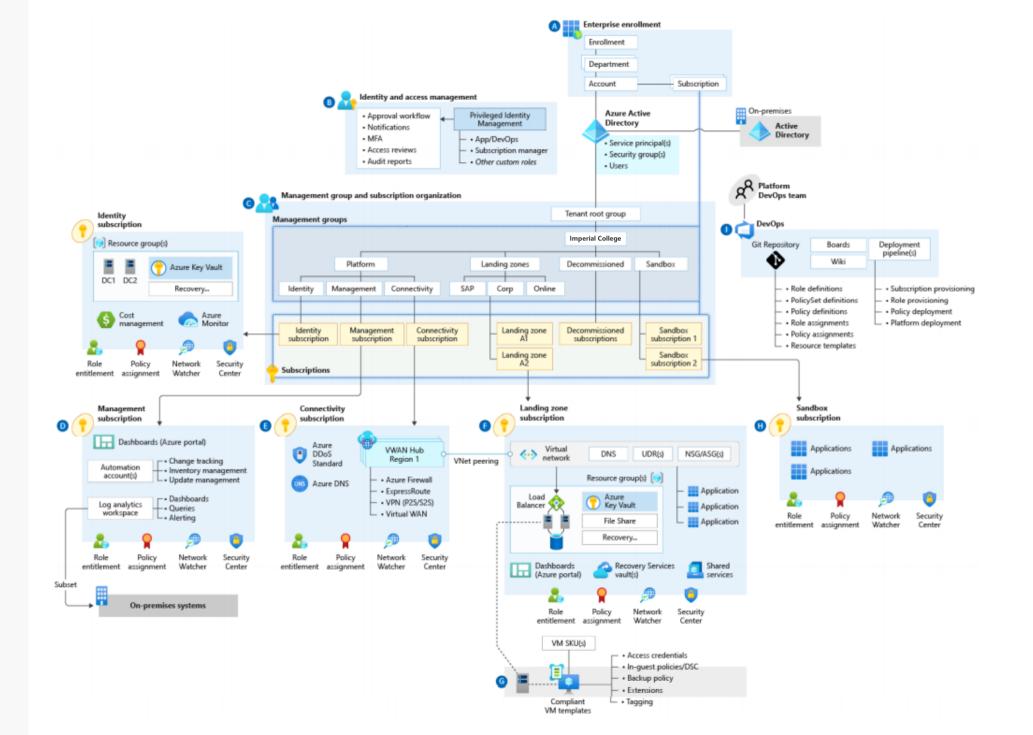


Azure Management Structure

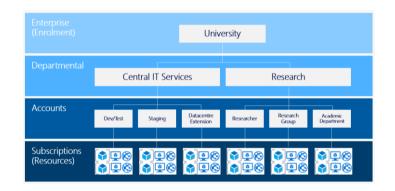


Azure Management Structure





Centrally managed Azure for Research



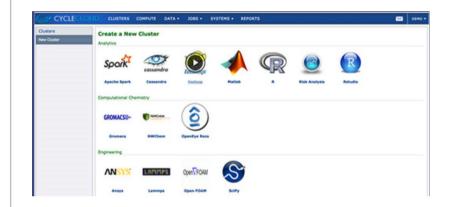
Central Research Management



Research structure, billing, account and delegation management. Central procurer Governance of Research services is a key aim of most Universities.



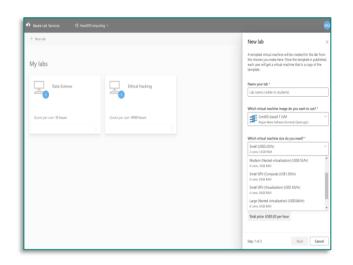
Immediate access to resources. No procurement process or pre-build saves time that can be directed to actual



On demand Research services

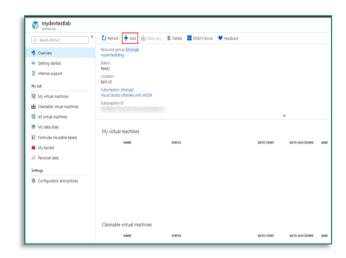
Researchers able to select their own resources without having to learn Azure or connect with Central IT with Azure Labs and CycleCloud

Azure resource management portals



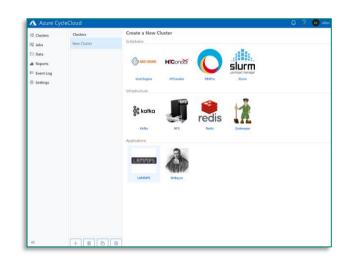
Azure Labs

- Curriculum use
- Lecturer manages
- Schedule sessions



Azure Dev/Test Labs

- Research development
- Central control



Cycle Cloud

- Manage clusters, HPC, HTC
- Scripted build
- Utilisation management
- Cost Management

HPC Resource Stack on Azure



Transformative Services

Azure Machine Learning

Azure Data Lake

Azure ML Compute



Workload Orchestration

VM Scale Sets

Azure Batch

Azure CycleCloud



Fast, Secure Networking

ExpressRoute

InfiniBand



High Performing Storage

Azure HPC Cache

Azure NetApp Files

Clusterstore



Optimized Compute

H-Series

N-Series

Cray

Machine Learning on Azure



Domain-specific pretrained models

Vision

Speech

Language

Search



Familiar data science tools

Visual Studio Code

Azure Notebooks

Jupyter

Command Line



Popular frameworks

PyTorch

TensorFlow

Scikit-Learn

ONNX



Productive services

Azure DataBricks

Azure Machine Learning

ML VMs

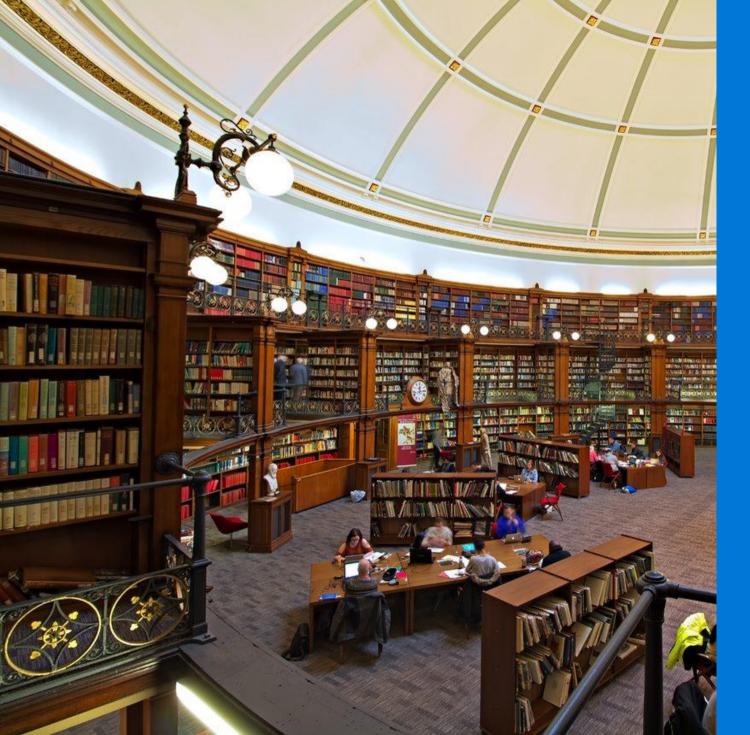


Powerful infrastructure

CPU

GPU

FPGA



HPC Resources on Azure

HPC VMs on Azure

No-compromise CPU and GPU based resources



- Up to 16 cores, 3.2 GHz E5-2667 V3 Haswell processor
- Up to 224 GB DDR4 memory, 14GB per core
- FDR InfiniBand @ 56 Gbps, 2.6 microsecond latency
- 2 TB of local SSD



- Up to 44 cores, Intel Xeon Platinum processor
- Up to 352 GB DDR4 memory, 8GB per core
- EDR InfiniBand @ 100 Gbps
- 700 GB NVMe



- Up to 60 cores, AMD EPYC processor
- Up to 240 GB DDR4 memory, 4GB per core
- EDR InfiniBand @ 100 Gbps
- 700 GB NVMe

H-Series:

Most powerful CPU virtual machines with optional RDMA



GPU virtual machines specialized for graphic-intensive workloads



- Up to 4 NVIDIA Tesla K80 GPUs
- Up to 24 cores
- Up to 224 GiB memory
- Up to 1440 GiB of local SSD
- FDR InfiniBand



- Up to 4 NVIDIA Pascal P100 GPUs
- Up to 24 cores
- Up to 448 GiB memory
- Up to 3 TB of local SSD
- FDR InfiniBand



- Up to 4 NVIDIA Volta V100 GPUs
- Up to 24 cores
- Up to 448 GiB memory
- Up to 3 TB of local SSD
- FDR InfiniBand



- Up to 4 NVIDIA Pascal P40 GPUs
- Up to 24 cores
- Up to 448 GiB memory
- Up to 3 TB of local SSD
- FDR InfiniBand



- Up to 4 NVIDIA Tesla M60 GPUs
- Up to 24 cores
- Up to 224 GiB memory
- Up to 1440 GiB of local SSD



- Up to 4 NVIDIA Tesla M60 GPUs
- Up to 24 cores
- Up to 448 GiB memory
- Up to 2,948 GiB of local SSD

Cray in Azure

A dedicated supercomputer on your virtual network



Innovate faster to solve your toughest challenges



Access to advanced cloud solutions and Al



Bring your simulation data and services closer together



Simplify the management of your infrastructure



Azure Batch Capabilities



Job scheduling

- Supports both embarrassingly parallel and tightly-coupled MPI jobs
- Run > 1 task in parallel per node
- Detect and retry failed tasks
- Can set max execution time for jobs and tasks
- Task dependencies
- Job prep and cleanup tasks



Rich app management

- Get apps from blobs, Batch app packages, package managers, custom VM images
- Docker container images



Choice of VMs

- Windows or Linux
- Standard or custom images
- Windows pool can use AHUB
- Use low-priority VMs



Monitoring

- VM monitoring and auto-recover
- Metrics and logs available via Portal and API

101010 010101 101010

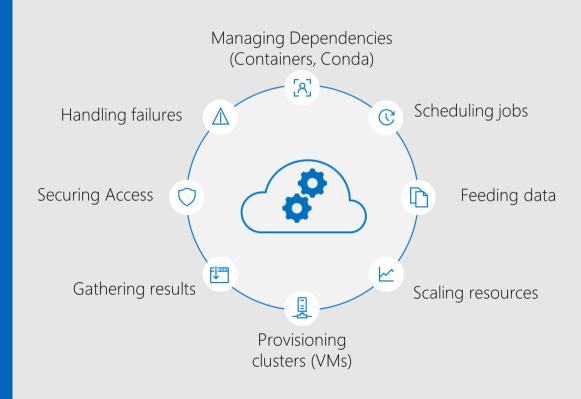
Access via APIs, CLIs, and UIs

- .NET, Java, Node.js, Python, REST
- PowerShell, x-plat Azure CLI
- Azure Portal, Batch Labs x-plat client UI

What is Azure ML Compute?

Addresses the challenges of training Machine Learning models

Challenges with ML Training . . .



... addressed by Azure ML Compute

- Integrates Batch AI to provide managed compute clusters and job scheduling
- Work with clusters of GPU or CPU to run experiments in parallel and at scale to reduce training time
- Enables data scientists to easily train, test, and score Deep Learning and other AI/ML models without managing infrastructure
- Use with Azure Machine Learning service and integrate with your data science tools and pipeline

Azure ML Compute Value



Dependencies and Containers

Leverage system-managed AML compute with pre-configured containers



Schedule jobs

Queue and prioritize jobs, start MPI, monitor and diagnose failures



Provision clusters

Use the latest NDv2 series VMs with the NVIDIA V100 GPUs. Install drivers and software, interact through SSH



Distribute data

Manage and scale access to training data, track logs & models in a secure & compliant manner



Scale resources

Autoscale resources to only pay while running a job. Share reserved instances, and use Low Priority VMs

What is Azure CycleCloud?



User Empowerment

- Cloud-enable existing workflows
- Enable instant access to resources
- Provide auto-scaling, error handling



IT Management

- Link workflows for internal and external clouds
- Use Active Directory for authentication and authorization
- Provide secure and consistent access

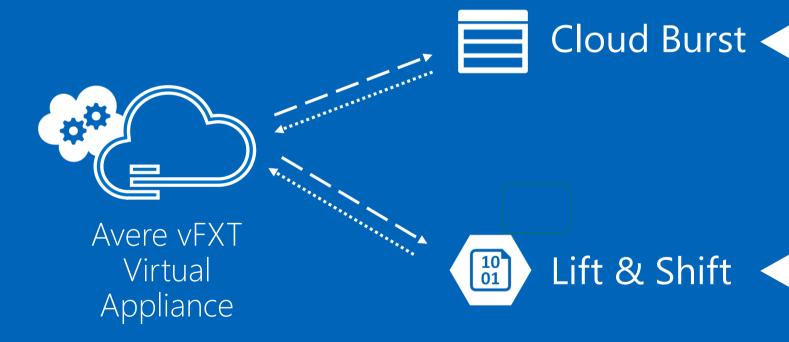


Business Management

- Increase accountability by directly coupling service usage to expenditures
- Provides the tools to better manage and control costs

What is Avere vFXT?

A high performance file caching system that bridges your HPC environment with Azure



On-demand expansion of your HPC environment in Azure, which supports growing and unpredictable business demands

A complete shift of your computing environment into Azure primarily for file-based HPC applications

Access Flexibility

Big Compute Evolution – Where Are You?

On-premises Clusters

- Easy scaling to reduce runtimes
- Job scheduling and management
- Compute node provisioning
- Hardware is paid for



Hybrid Clusters

- Extend cluster to cloud to handle peak demand
- CycleCloud and Azure
- Manage like machines on your network with VPN or Express Route



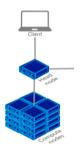
IaaS Clusters

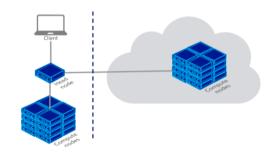
- Deploy cluster all in cloud
- Move existing applications
- Support projects & testing
- Gallery images and templates to deploy
- Flexible VM configuration

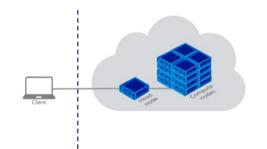


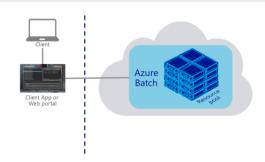
Cloud Native

- Azure Batch: Native cloud scheduler
- DevOps, not infrastructure management
- Elasticity with auto-scale
- Small to very large deployments
- Use within a service or to offer SaaS



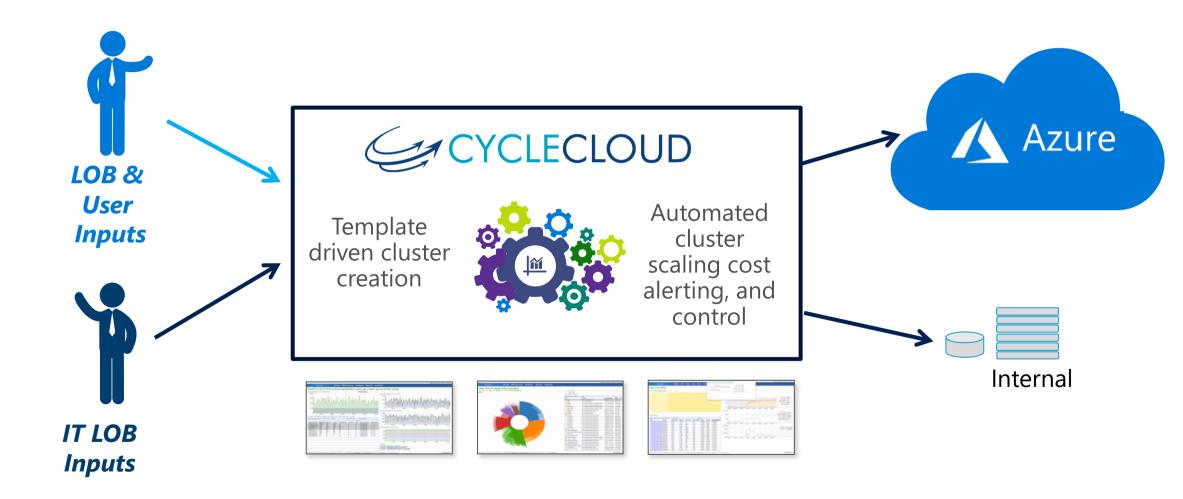








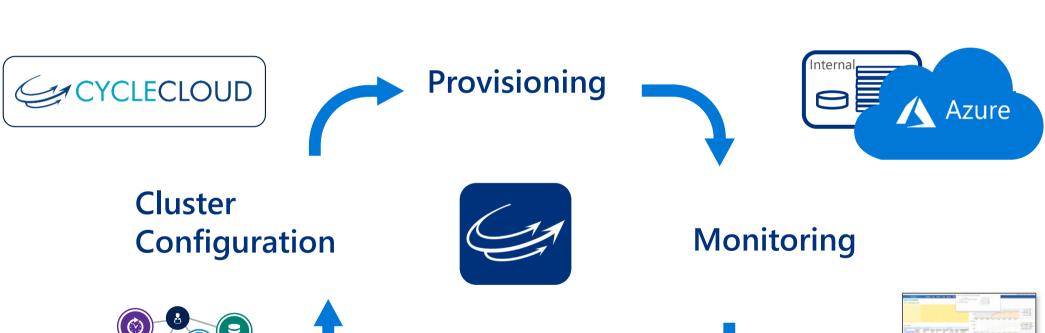
Making Users and IT Successful





Hybrid/Clustered Big Compute Lifecycle









Optimization







When to Use Azure Batch or CycleCloud

CycleCloud – Running clusters

- Targeted for IT/LOB Big Compute users running clusters
- Goal is to support hybrid workloads or clustered applications
- Clone existing clusters/workflows
- Support for HPC style clusters
 (PBSPro, Data Synapse, SLURM, GridEngine)
 and BigData/ML (Spark, TensorFlow)
- Azure-enable existing apps without rewriting

Azure Batch – Running jobs

- Targeted for developers writing applications that submit jobs
- Goal is to run jobs, with no cluster framework
- Accelerators for rendering, Al training, etc.
- No need for clusters/schedulers
- Enables creation of "SaaS" type services



Unleashing Big Compute Productivity





Users can cloud-enable existing workflows
Instant access to resources
Auto-scaling, error-handling



Reliable tools for IT

Workflows linking internal and external clouds

Tools to manage & control costs



Organizations providing secure, consistent access
Ability to link usage to spend
Leverage AD for authorization/governance

Microsoft Confidential

Complete Multi-Cloud Workflow Control



Scope

Configure

Run on Cloud

Optimize

Internal



amazon web services

Cloud Platform

■ Windows

Azure

Google





- Workload placement
- Optimal scale
- Cost optimization
- Data scheduling



Reporting



Cluster Configuration

- Multi-cloud, without changes
- Pre-set or User-defined "types"
- Abstraction for all cluster data. attributes (roles, OS, etc.)



File: Declarative Cluster Definition



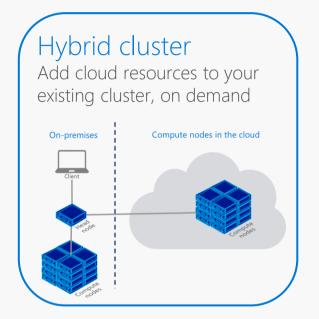
Packages, Installers Containers, Data

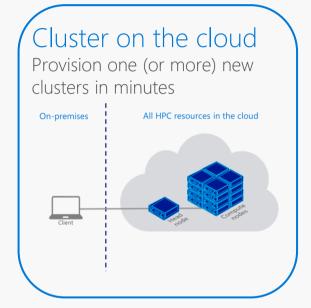
Optimization

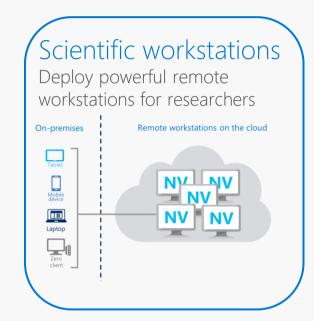
- Benchmark instances
- Make Workflow UI
- Human Workflow

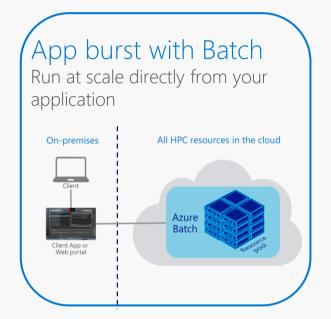


Azure HPC Solutions







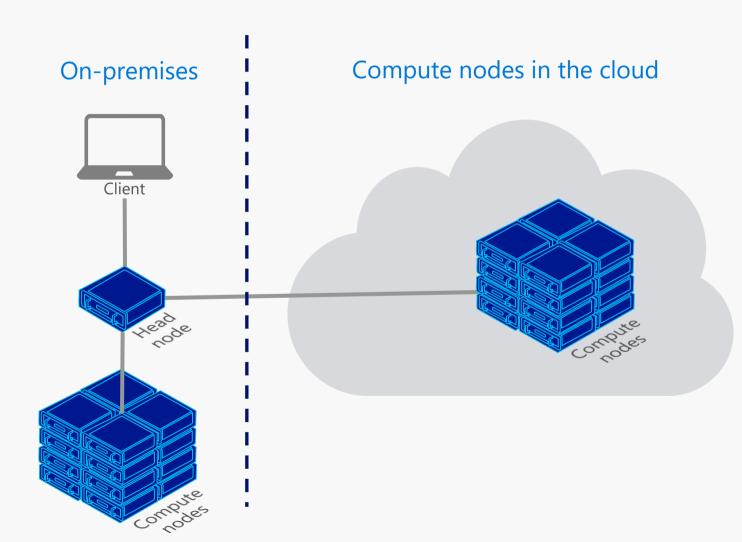


	HPC resources on demand		App Burst	Remote Scientific Desktop	
Target persona vs. Solution	Hybrid burst	laaS cluster	Batch	GPU workstations	
Individual researcher/scientist		✓	✓	✓	
Director of research group	✓	✓	✓	✓	
Research IT team lead	✓	✓		✓	
Central IT group lead	✓		✓	✓	

Software & services: Cloud burst

Add cloud resources to your cluster, on demand

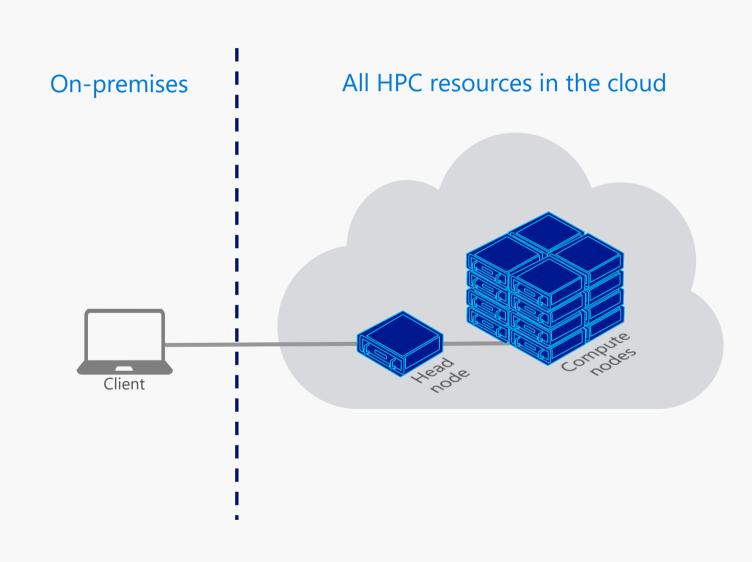
- Burst to cloud to add resources when needed
- Continue using your existing on-premises infrastructure
- Run workloads on Windows and Linux, on Azure and on-premises
- Cover peaks in demand or special projects
- Pay only for what you use
- Microsoft HPC Pack, <u>Univa Grid Engine</u>, and <u>Altair PBS Pro</u> already have this capability (and more are coming!)



Software & services: Cluster on the cloud

Provision one (or more) new clusters in minutes

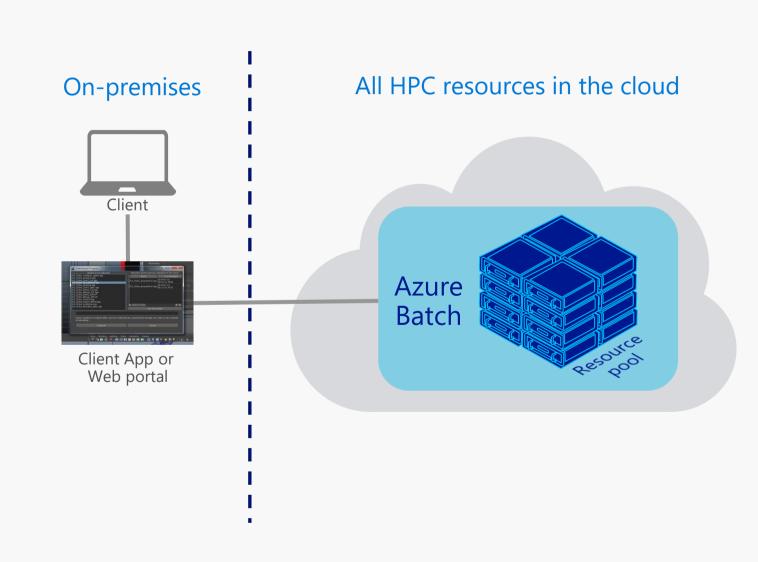
- Deploy a complete HPC cluster, all in the cloud, in just minutes
- Use templates, scripts, and images to quickly deploy at any scale
- Use your current HPC scheduler
- Shift existing applications, scripts and tools to cloud
- Deploy as many clusters as you need!



Software & services: HPC as a service

Run at scale directly from your application

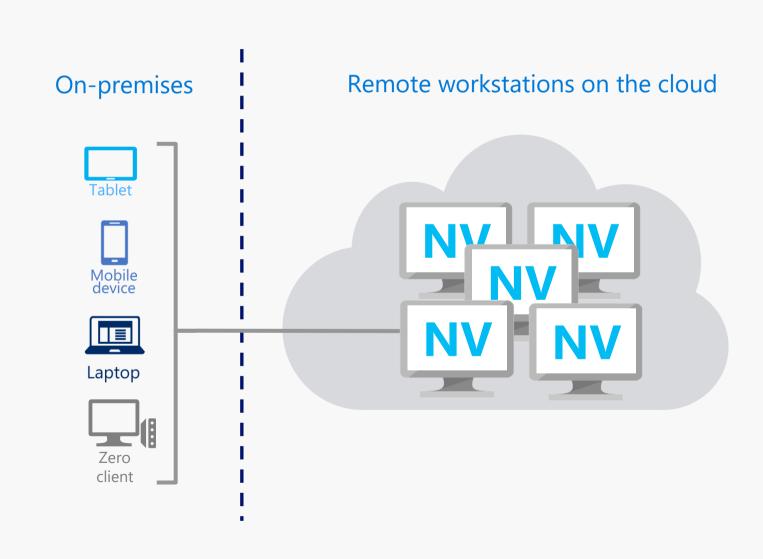
- Integrate with Azure Batch, directly from a client application (GUI or CLI) or online portal
- Batch abstracts resource management and scheduling completely
- Supports small to extremely large deployments and can deploy any VM size
- Provides auto-scaling and stopping of resources
- Run HPC jobs at scale on Docker containers
- Using Batch is free, you only pay for the underlying compute!



Remote Scientific Workstations

Deploy powerful GPUenabled workstations

- Deploy one or more GPU-enabled workstations on the cloud
- Connect from any laptop or device (iOS, Android, Windows)
- Collaborate, share, explore.
- With or without optimized remote visualization platforms (e.g., Teradici PCoIP or Citrix HDX)

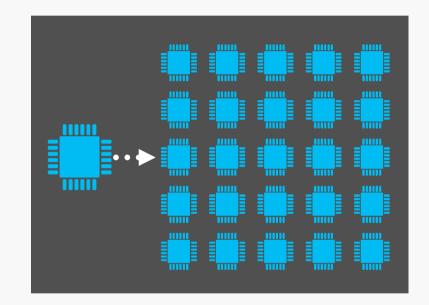


What would you do with 100x the scale?

Do more with hyper-scale:

- Service more users
- Run more projects
- Get results faster

- Run larger simulations
- Explore new insights (e.g., "What if?")



Remove current limitations:

- Modify more parameters
- Analyze more complex models
- Visualize larger results

- Run more iterations
- Generate higher fidelity results
- Simulate longer periods of time

Breakthroughs Enabling Intelligence



Microsoft Cognitive Services

Give your apps a human side













Vision

From faces to feelings, allow your apps to understand images and video



Hear and speak to your users by filtering noise, identifying speakers, and understanding intent

Language

Process text and learn how to recognize what users want

Knowledge

Tap into rich knowledge amassed from the web, academia, or your own data

Search

Access billions of web pages, images, videos, and news with the power of Bing APIs

Labs

An early look at emerging Cognitive Services technologies: discover, try & give feedback on new technologies before general availability

Microsoft Genomics

Power genome sequencing and research insights

- Cloud implementation of Burrows-Wheeler Aligner (BWA) and the Genome Analysis Toolkit (GATK) for secondary analysis
- ✓ ISO-certified and covered by Microsoft under the HIPAA BAA
- ✓ Global coverage to meet your data sovereignty needs
- ✓ Only pay for what you use with no commitment
- ✓ Easy-to-use API for integration with partner solutions

Strategic help from Microsoft

University Vision, Strategy and Business Objectives

Methodology, Governance and planning

Assessment, readiness and cost estimation

Architecture & Design, Deployment and training

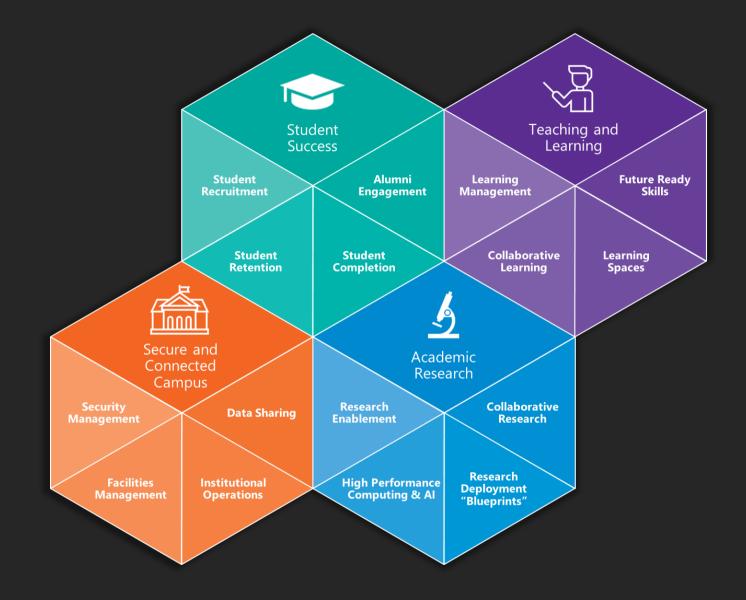


Cloud Adoption Framework

Cloud Economics Assessment

Azure Migration Program

Align to
University strategy and
objectives



Strategic Engagement



Empowering Transformation









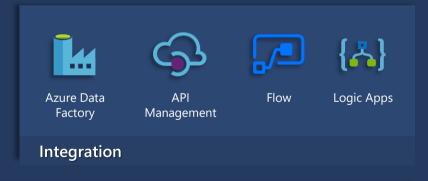
Identity















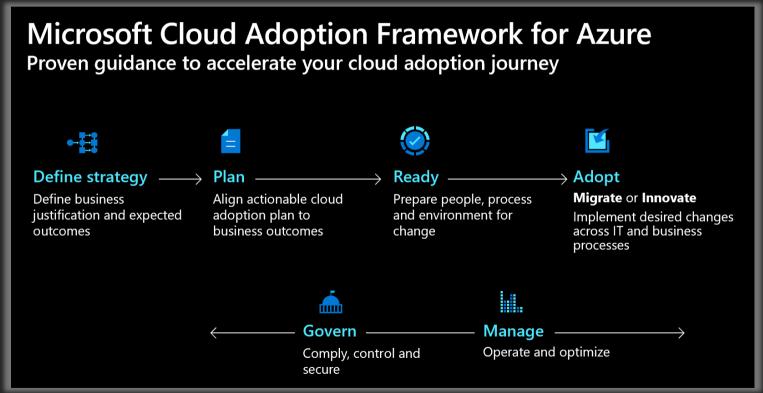


Align to
University strategy and
objectives

Methodology, Planning and Governance



Education Transformation Framework



Strategic Cloud Engagement

Tools, templates, and assessments



Cloud journey tracker

Business outcome template



Azure DevOps demo generator

Cloud adoption plan template



Azure setup guide

Readiness checklist

Naming and tagging tracking template

Landing zone blueprints



ιαορι

Azure migration guide

Strategic migration assessment

and readiness tool (SMART)

Azure innovation guide



Govern

Governance benchmark

Governance process template

Cost Management process template

Deployment acceleration process template

<u>Identity process template</u>

Resource consistency process template

Security baseline process template



Manage

Microsoft Azure Well-Architected Review

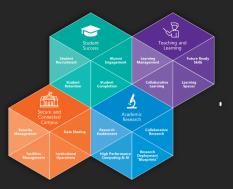
Best practices source code

Operations management workbook

Align to
University strategy and
objectives

Methodology, Planning and Governance

Assessment, readiness and cost estimation







Education Transformation Framework

Cloud Adoption Framework

Cloud Economics Assessment

Strategic Cloud Engagement

Carbon negative by 2030 - erase footprint by 2050



Ground our work in the best available science and the most accurate math



Take
responsibility
for Microsoft's
own carbon
footprint, getting
to negative by
2030, 55% Scope
3 reduction



Fund \$1B for better carbon reduction and removal



Support and empower suppliers and all customers with new tools, products and partnerships



Work to advance transparency for reporting on emissions and removals



Use our voice on carbon-related public policy issues



Enlisting our employees to enable them to contribute to our efforts







What it means for customers



Ground our work in the best available science and the most accurate math



Take
responsibility
for Microsoft's
own carbon
footprint, getting
egative by
55% Scope
ction



Fund \$1B for better carbon reduction and removal



Support and empower suppliers and all customers with new tools, products and partnerships



Work to advance transparency for reporting on emissions and removals





Use our voice on carbon-related public policy issues



Enlisting our employees to enable them to contribute to our efforts



Microsoft cloud services are up to 98% more CO₂ efficient vs. on-prem Microsoft cloud services are up to 93% more energy efficient vs. on-prem

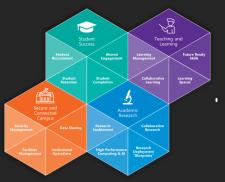
All Surface devices meet or exceed US and EU green IT standards New tools available to developers and organizations

Align to
University strategy and
objectives

Methodology, Planning and Governance

Assessment, readiness and cost estimation

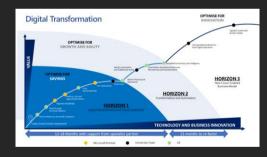
Architecture, Design, Deployment and training



Microsoft Cloud Adoption Framework for Azure
Proven guidance to accelerate your cloud adoption journey

Pofine strategy
Define strategy
Migrate or Innovate
Implement desired charges
across IT and business
processes

Manage
Operate and optimize
secure



Education Transformation Framework

Cloud Adoption Framework

Cloud Economics Assessment

Azure Migration Program

Strategic Cloud Engagement



Azure Migration Program (AMP)

Simplify your path to the cloud













Proven methodology: Cloud Adoption Framework for Azure Cost-effective offers & incentives

Technical skill building

Free cloud migration tools: Azure Migrate

In-depth assistance from FastTrack Azure engineers Expert help from specialized partners

Your cloud journey

Pre-migration

Migration

Post-migration



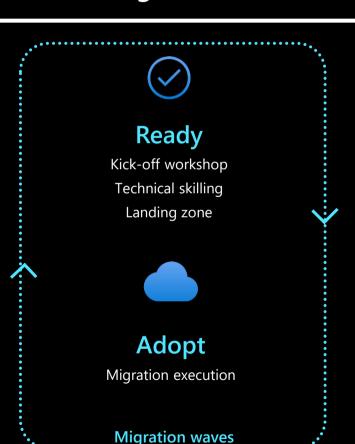
Define strategy

Executive sponsorship Stakeholder alignment Partner engaged



Plan

Discovery and assessment TCO/ Business case
Migration plan





In Summary



Azure for Research

- Central IT Service
- Governance built in
- Secure and Compliant
- Faster to research publications
- Virtually immediate access
- Scalable to demand
- Kickstart initiative



Thank you

