ANSYS Cloud Architecture and Security Overview

Introduction

ANSYS Cloud is a multi-tenant PaaS service in which a single instance of a software application serves multiple customers. Each customer is called a tenant, each of whom is part of a group of users who share a common access with specific privileges to the storage and software instance. This paper covers the following topics:

- Regions and Geographies
- Data Retention and Deletion Policy
- Scalability
- Threat Modeling
- Data residency
- Functionality
- Support
- Architecture

Azure Regions and Geographies

Azure is organized into regions and geographies (Geos). A region is a set of data centers deployed within a latency-defined perimeter and connected through a dedicated, regional low-latency network. A geography is a discrete market, typically containing two or more regions, that preserves data residency and compliance boundaries.

Geographies allow customers with specific data-residency and compliance needs to keep their data and applications close. Geographies are fault-tolerant to withstand complete region failure through their connection to Azure's dedicated high-capacity networking infrastructure. Furthermore, each Azure region is paired with another region within the same geography, together making a regional pair. Across the region pairs Azure will serialize platform updates (planned maintenance) so that only one paired region will be updated at a time. In addition, in the event of an outage affecting multiple regions, at least one region in each pair will be prioritized for recovery.

ANSYS Cloud Service is architected to take full advantage of this feature by deploying shared resources in fault-tolerant, high-availability region pairs.

Data Residency

ANSYS Cloud will not store customer data outside the customer-specified Geo except for the profile information that is stored in Azure Active Directory B2C, which is based on the geography of the company/tenant location. In this case, the ANSYS tenant is in the United States. All customer simulation data is stored in Azure Globally Redundant Storage (GRS) and the data is copied between two regions within the same Geo for enhanced data durability in case of a major data center disaster. All virtual machines used for compute are started in the customer-specified Geo, so data residency persists based on the selected region.

Data Retention and Deletion Policy

ANSYS Cloud will not delete any data on a customer's behalf, and therefore there is no data retention and deletion requirement. Each customer is solely responsible for its data retention and deletion policy and procedure. ANSYS will invoice each customer based on the size of the data residing in storage.

Functionality

The ANSYS Cloud Service application is deployed in multiple regions to achieve high availability. It focuses on an active/passive scenario with hot standby, using Traffic Manager for failovers. The architecture consists of the following elements:

- Primary and secondary regions. This architecture uses two paired regions (East U.S. and West U.S., for example, when deployed in the U.S. Geo) to achieve higher availability. The ANSYS Cloud service application is deployed to each region. During normal operations, network traffic is routed to the primary region. If the primary region becomes unavailable, traffic is routed to the secondary region.
- Azure Traffic Manager. Traffic Manager routes incoming requests to the primary region. If the application running that region becomes unavailable, Traffic Manager fails over to the secondary region.
- Geo-replication of Cosmos DB. One region is designated as writable and the other is a read-only replica. If there is a regional outage, an automatic failover will take place which selects the other region to be the write region. The client SDK automatically sends write requests to the current write region, so there is no need to update the client configuration after a failover.
- Resource groups. Deployment is done by placing the primary region, secondary region and Traffic Manager into separate resource groups. This lets us manage the resources deployed to each region as a single collection.
- Isolated Compute regions. In addition to deploying shared resources in a highly redundant way, each compute region consists of only enough resources needed to successfully execute a job. It is deployed into numerous regions and a customer is offered a choice of which region to use for running simulation jobs. All customer data persist in the customer-specified region and are never copied outside of the Geo. Each compute region consists of the following resources:
- VNet, Subnets and Security. Each region has its own VNet defined with unique CIDR and IP address allocation.
- Batch Service. Each region has its own Batch Service account with access to custom-made images with ANSYS solvers pre-installed and configured.
- Storage Accounts. Each region has its own storage accounts provisioned so that simulation job data reside in the same data center as the compute nodes.

Scalability

ANSYS Cloud Service utilizes Azure PaaS Web Applications service. Each service plan is placed in an auto-scaling group, which mitigates unexpected high bursts in usage. The applications do not store any data; therefore, in case of service interruption no data is lost. In case of a hardware failure Azure will restart the service automatically on another hardware system. Finally, in case of a region outage our secondary backup region should take over and continue to serve the application.

Threat Modeling

Microsoft and ANSYS conducted a threat modeling exercise for the ANSYS Cloud product. The goal was to determine if the application uses Azure and product defenses correctly and to ascertain any potential security weaknesses.

Two documents were created to support this effort: an initial free-form Word document using a threat model template and a threat model built in the Microsoft Threat Modeling Tool using the SDL Knowledge Framework Azure template. Below are the key points arising from the threat model process.

- One of the benefits of hosting a system that comprises mainly Azure PaaS components is that many of the security controls are addressed by Microsoft. For example, anti-malware and system security updates are both addressed automatically by Microsoft[1].
- Azure provides network denial of service protections, reducing the potential for availability disruptions.
- Local redundancy is built in[2], and Geo-redundancy[3] is also available, as needed, for Azure solutions.
- The system is segmented using Network Security Groups and Virtual Networks to ensure confidentiality and integrity.
- TLS 1.2[4] is used by default for all network traffic, encrypting all transmissions of information.
- Cryptography is implemented in the file upload functionality to ensure the security and protection of data every step of the way.

[1] Access the Azure compliance documentation to read more about the external Azure audit findings on these topics.

[2] Files are written to multiple locations at the same time.

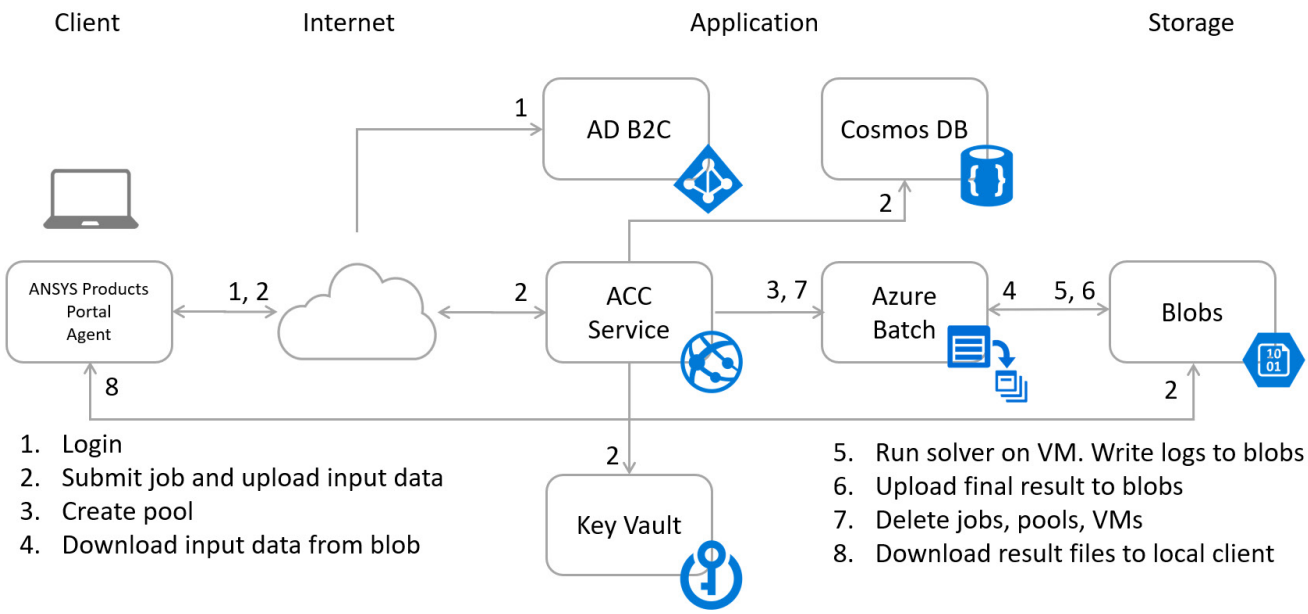
[3] Fulfilled through Cosmos DB – deployed into two regions (active and passive).

[4] Enforced by Azure PaaS and SaaS services.

Architecture

The architecture of ANSYS Cloud is best described through a job submission sample. The application can largely be group into three distinct tiers:

1. Client. This layer comprises a web-based portal and ANSYS Cloud Extension agent, which is integrated with the ANSYS desktop application.
2. Application. The Application tier consists of a front-end layer (ANSYS Cloud Service) and a back-end layer (Azure Batch). The front-end tier is responsible for authentication and for processing input parameters in order to define a workflow to be processed by the back-end tier.
3. Storage. The back-end tier will download input data files from Azure storage, process the models as specified by the workflow and upon completion upload the output data to Azure storage.



Security is built-in to ANSYS Cloud from the ground up. The different steps are be described below.

1. Login

ANSYS Cloud is registered with AD B2C, which supports the OpenID Connect and OAuth 2.0 standard protocols. Customers requesting access to ANSYS Cloud will be redirected to AD B2C for authentication. Upon successful verification AD B2C issues a JWT token to the web browser, which is redirected back to ANSYS Cloud. The ANSYS Cloud web service validates the JWT token and the user is signed in to the application. All communication between the web client, AD B2C and the ANSYS Cloud service is protected by HTTPS/TLS.

2. Submit Job and Upload Data

All data, both in motion and at rest, are encrypted. The key pair is fully managed by ANSYS for the end user by storing the public and private key in the Azure Key Vault; only the public key is shared. To ensure data confidentiality, all simulation data are encrypted using a symmetric key pair (AES 256 based) which is downloaded via an authenticated REST API call over HTTPS/TLS to our ANSYS Cloud back-end service, which are then encrypted via asymmetric key.

The encrypted data files are uploaded to a customer specific storage container (ACL) in Azure Blob storage using the Azure Storage SDK. In addition, Shared Access Signature (SAS) tokens are generated using a strict permission policy, including the start time and the expiry time.

3. Create Pool

The pool is the fundamental security perimeter in Azure Batch. Jobs running in different pools do not have access to any VMs or data used by a job in a different pool. In ANSYS Cloud each job runs in a dedicated pool. Virtual machines created for a job communicate over a subnet that is reserved exclusively for a particular customer. In addition, the network is protected by a Network Security Group, restricting access from the Internet and from other Azure networks.

4. Download Data from Azure Storage to VM

As part of the 'job submit' stage a workflow is defined by the ANSYS Cloud service. The first step in this workflow is to download and decrypt data from Azure storage and store the decrypted data on the temporary SSD disk of the customer's dedicated VM. All communication between the Azure compute and storage tiers is protected by HTTPS endpoints. Upon Batch Pool creation, a temporary certificate is generated for each run and associated with the pool that is used to encrypt each symmetric key for the solve. When an ANSYS Cloud service needs to access a simulation data file, the service gets the decrypted symmetric key using a certificate managed by the ANSYS Cloud service. Since the certificate is tied to a pool, once the job is done all resources associated with that pool are deleted along with temporary certificates.

5. Run Solver on VM

The main step of ANSYS Cloud is the actual MPI execution of the computation binaries (Fluent/Mechanical/HFSS). The Cloud Compute workflow consists of multiple data types categorized as follows:

1. **Simulation Input Data.** (Already downloaded and decrypted from the Azure storage in the previous step.) Model input data includes all the required modeling data to run the simulation. This includes:
    1. Solver input file. A script file generated by the ANSYS applications (pre-processing phase) that describes the simulation model. This file is based on the language compatible with the solver to be used (MAPDL for Mechanical, SCHEME for Fluent).
    2. User subroutines (used for solver customization). These contain additional source files that have been created by the customer to be consumed by the solver during run time (custom made libraries).
    3. User macros (used for solver customization). These contain additional script files that have been created by the customer to be consumed by the solver during run time.

2. **Simulation Output Data.** (Encrypted and uploaded to customer-specific container with ACL controls in place.) Contains all data generated by the simulation process and persisted in solver output files. The output data varies for each ANSYS solver; you can find more detailed information about the outputs for the different solvers on the ANSYS help website. However, the items below describe the three main categories of output we expect from the solvers:
    1. Solver result files. These files contain all physical quantities computed by the solver.
    2. Solver log file. This file contains all information related to the progress of the computation. It details the solver convergence and hardware consumption (CPU and memory allocation).
    3. Custom output files. Contains all data generated by user subroutines and macros as described in the Simulation Input Data. This can be additional results or information about the computation.

3. **Monitoring Data.** Monitoring is performed at different stages in the cloud compute workflow, and we collect different information at these stages. Workflow monitoring data is all the data that we collect from the workflow engine service, and there is no link with any data provided by the customer. This service centralizes the way the job is submitted on the cloud and provides information on the process, the resources and the progress. In addition, we collect information on the user who launched the job. All data is stored in our back-end database system for easier tracking.

Below is the specific Workflow Monitoring Data that is captured.

1. Workflow state: start/stop/failed
2. Task state: start/complete: a workflow is decomposed in steps and we can get a status for each of the steps
3. Transfer rates: for both upload and download, these rates provide information about the performance of the file transfers
4. Exceptions (error message): all exceptions that the workflow engine services generate
5. Simulation run times: (start time/end time)
6. Upload/Download times: (start time/end time)
7. Time to create VM pools: (start time/end time)
8. Region in which the simulation was executed
9. Configuration (type of VM, number of VM)
10. Credits
11. Session: The session is created when the job is submitted. Any action part of the job submission is integrated in the session, which allows a clear isolation of the job.

6. Upload Results to Blob Storage

Upon termination of the MPI computation, the result files are encrypted on the VM using the public key and a self-generated symmetric key using the certificate attached to the pool. Blobs are encrypted and the encryption key, alongside the RSA key, is added as metadata to the blob and uploaded to blob storage in a user-specific container.

7. Delete Jobs, Pools and Virtual Machines

All resources (jobs, pools and VMs) created for the job will be deleted upon job termination. Once a VM is restarted or deleted there is no way to recover the data on the temporary drive .

8a. Download for Local Post-Processing

When a user requests to download the simulation data to a local workstation for further post-processing, an encrypted file is retrieved from the blob storage first over HTTPS/TLS. For each blob, the agent that runs on the client desktop is making a call to the ANSYS Cloud back-end service via an authenticated REST API call over HTTPS/TLS to get the decrypted symmetric key. Additionally, a check is done to verify that the signed-in user is indeed accessing and has enough permission for the file in question. Finally, the file is decrypted locally using the symmetric key.

8b. Download for Post-Processing in Azure

Upon receiving a post-processing request, a pool creation is initiated and a remote visualization resource is allocated. Upon Batch Pool creation a temporary certificate is generated for each run and associated with the pool. Since the post-processing resource is behind an NSG with tightly controlled access rules, a remove visualization resource endpoint and credentials are registered in the publicly facing gateway, which acts as a reverse proxy processing all clients' requests to the upstream services over HTTPS/TLS.

When an ANSYS Cloud service needs to access a simulation data file, as for Azure Batch service, the post-processing service uses a certificate to decrypt the file. Since the certificate is tied to a pool, once the job is done all resources associated with that pool are deleted along with temporary certificates and any registered services in the gateway system.

Additional Resources: Microsoft Trust Center