AppScale ATS

Operational Brief

This document defines the relevant terms and describes, at a high level, deployment configurations, operational situations, and operator's activities for AppScale ATS.

Subsystems

AppScale ATS deployments consist of a variety of components potentially deployed across many hardware nodes, with each component serving a specific role. It is useful, however, to group components into just four subsystems:

- **Control** - Stateful components with a cloud-wide mandate.
- **AZ Control** - Stateless components responsible for a specific Availability Zone.
- **Storage** - Stateful components implementing backing for object and block storage.
- **Compute** - Stateless components enabling execution of virtual machines.

Configurations

AppScale ATS can be deployed in four standard configuration types. Converting from one type to another is not recommended as it involves down-time and manual migration process.

1. **Extra Small (XS)** - *Non-durable, one-node deployment.* This is a single-node deployment that cannot be expanded, intended for proof-of-concept setups to demonstrate the features of ATS. Control, AZ Control, Storage, and Compute subsystems are all co-located. All features of ATS are available, but performance and capacity of the system is limited. Failure of the node may result in total loss of data.*

2. **Small (S)** - *Non-durable, minimally scalable deployment on 2+ nodes.* This is a deployment with one control-plane node (co-locating Control, AZ Control, and Storage subsystems) and one or more Compute nodes. This configuration is intended

for proofs-of-concept or for small setups (e.g., test/dev rigs) for which data loss is acceptable. This configuration will not scale far in terms of compute, network, and storage performance or capacity. Failure of the controller node may result in total loss of data.☘

3. **Medium (M)** - *Durable hyper-converged deployment.* Deployment with two or more control-plane nodes (at least one Control node and one AZ Control node) and three or more Compute nodes that double as Storage nodes. In this configuration compute and storage capabilities scale in tandem, allowing cost-effective deployments of a range of sizes. There is redundancy for outside connectivity and for data (block and object). These deployments are limited to one AZ.

4. **Large (L)** - *Durable dedicated-storage deployment.* Deployment with two or more control-plane nodes (one Control node and one or more AZ Control nodes), three or more dedicated Storage nodes, and one or more dedicated Compute nodes. In this configuration compute and storage capabilities can be scaled separately, at the additional cost of dedicated Storage nodes. There is redundancy for outside connectivity and for data (block and object).

☘ - unless external persistent volume is utilized

The configuration types have the following characteristics:

Type	Nodes	Storage	Networking	Console	Use cases
XS	1	Linux file system	VPC	yes	POC
S	2 - 9	Linux file system	VPC	yes	POC, dev/test
M	5 - 40	Ceph on compute	dual-gateway VPC	yes	one-AZ production workloads
L	6 - 100's	Ceph dedicated	dual-gateway VPC	yes	multi-AZ production workloads

Failure scenarios

AppScale ATS was designed to keep the data and control planes separate, thus ensuring that cloud workloads continue to operate – with network and storage connectivity – even when some control-plane components are temporarily unavailable.

The following table considers the implications of a fail-stop failure (either a **transient** one, solvable with a reboot, or a **permanent** one, such as a storage device failure) for nodes deployed in configurations S, M, and L:

Failed subsystem	Failure effects	Recovery process
Control	API endpoints are unreachable (e.g., S3 is inaccessible, instances cannot be started, volumes cannot be attached). S-configurations only: Instances will lose network connectivity and EBS volume attachments. With permanent failure, EBS and S3 data will be lost, along with instances themselves.	*Transient failure:* Restart the node. *Permanent failure:* Add a new Control node with the same IP, restore cloud DB and cloud credentials from backup or from an external persistent volume.
AZ Control	Loss of control over the specific AZ (e.g., no instance or volume operations). S-configurations: See "Control" above.	*Transient failure:* Restart the node. *Permanent failure:* Add a new AZ Control node with the same IP and the cloud credentials. It will be automatically reintegrated.
Storage	M- & L-configurations: possible performance degradation, but no loss of data or functionality (up to the chosen redundancy level). S-configurations: See "Control" above.	*Transient failure:* Restart the node. *Permanent failure:* Add a new Storage node with same IP and the cloud credentials, register with Ceph, and wait for automatic rebalancing to complete.
Compute	Instances on the node will be lost. (With Auto Scaling, replacement instances may be started on other nodes.) Ephemeral data will be lost, but data in EBS volumes will persist.	*Transient failure:* Restart the node. *Permanent failure:* If it is desired to restore cloud compute capacity, add a new Compute node, give it the cloud credentials, and register with the cloud.

Adding a new node of a certain type – as specified in the recovery procedures above – implies booting up a server that has the appropriate software installed (operating system and AppScale ATS components and their dependencies), either via an image or an installation routine. Once cloud-specific credentials are added, the node is automatically integrated into the deployment.

Transient software failures, such as crashes of individual software components, are not enumerated above. Some software components restart automatically (e.g., Compute controllers), others may need to be restarted explicitly. Requests that were in flight at the time of the crash and up until the component is restarted will need to be retried by the user.

Node failure in an XS-configuration results in total loss of data and functionality, requiring a new XS deployment.

Operational Tasks

Nodes: adding compute

To add a compute node to an existing AZ, a physical node will need to be added to the network of the existing compute nodes, with the appropriate software installed. After starting the ATS component a registration from the AZ controller (using the `clusteradmin-register-nodes` command) will make the new node available.

Nodes: decommissioning compute

To decommission a running compute node, the `euserv-migrate-instances` command will move instances to other nodes. Then the node can be taken offline for maintenance or permanently. Cloud administrator can use the same command to migrate individual VMs.

Nodes: adding an AZ

To add a new AZ Controller to an L-configuration, a physical node with the appropriate software will need to be deployed. It may reside on a separate private network from other AZ controllers, but would have to be accessible by the Controller subsystem. Once the ATS components have been started, the `euserv-register-service` can be used to configure the new AZ Controller.

Accounts: creating

The account created during deployment is the account of the cloud administrator. Any user added to this account will be a cloud administrator and will have to ability to see workloads of all cloud users.

A cloud administrator can create ordinary cloud users with the `euca-create-account` command. The user thus created works as the account administrator and will have full power over the account, thus she can create new users, groups, and profiles within the realm of the account. The `euare-useraddloginprofile` command allows user to access the ATS console.

Backup & restore

The Controller subsystem stores on disk all the data that's pertinent to the deployment (users, groups, IAM policies, nodes, instances, volumes, etc.). That data are stored in a local database and the backup strategy requires capturing the database content. Restoring a failed Controller requires the restore of the database, either from a snapshot or from a persistent volume external to the deployment.

Backing up data held in a Ceph deployment (used as backing for S3 buckets and EBS volumes) is beyond the scope of this document.