The Exchange 2016 Preferred Architecture

The Preferred Architecture (PA) is the Exchange Engineering Team's best practice recommendation for what we believe is the optimum deployment architecture for Exchange 2016, and one that is very similar to what we deploy in Office 365.

While Exchange 2016 offers a wide variety of architectural choices for on-premises deployments, the architecture discussed below is our most scrutinized one ever. While there are other supported deployment architectures, they are not recommended.

The PA is designed with several business requirements in mind, such as the requirement that the architecture be able to:

- Include both high availability within the datacenter, and site resilience between datacenters
- Support multiple copies of each database, thereby allowing for quick activation
- Reduce the cost of the messaging infrastructure
- Increase availability by optimizing around failure domains and reducing complexity

The specific prescriptive nature of the PA means of course that not every customer will be able to deploy it (for example, customers without multiple datacenters). And some of our customers have different business requirements or other needs which necessitate a different architecture. If you fall into those categories, and you want to deploy Exchange on-premises, there are still advantages to adhering as closely as possible to the PA, and deviate only where your requirements widely differ. Alternatively, you can consider Office 365 where you can take advantage of the PA without having to deploy or manage servers. The PA removes complexity and redundancy where necessary to drive the architecture to a predictable recovery model: when a failure occurs, another copy of the affected database is activated. The PA is divided into four areas of focus:

1. Namespace design
2. Datacenter design
3. Server design
4. DAG design

Namespace Design

In the Namespace Planning and Load Balancing Principles articles, I outlined the various configuration choices that are available with Exchange 2016. For the namespace, the choices are to either deploy a bound namespace (having a preference for the users to operate out of a specific datacenter) or an unbound namespace (having the users connect to any datacenter without preference). The recommended approach is to utilize the unbounded model, deploying a single Exchange namespace per client protocol for the site resilient datacenter pair (where each datacenter is assumed to represent its own Active Directory site - see more details on that below). For example:

- autodiscover.contoso.com
- For HTTP clients: mail.contoso.com
- For IMAP clients: imap.contoso.com
- For SMTP clients: smtp.contoso.com

Each Exchange namespace is load balanced across both datacenters in a layer 7 configuration that does not leverage session affinity, resulting in fifty percent of traffic being proxied between datacenters. Traffic is equally distributed across the datacenters in the site resilient pair, via round robin DNS, geo-DNS, or other similar solutions. From our perspective, the simpler solution is the least complex and easier to manage, so our recommendation is to leverage round robin DNS. For the Office Online Server farm, a namespace is deployed per datacenter, with the load balancer utilizing layer 7, maintaining session affinity using cookie based persistence. pafig1 Figure 1: Namespace Design in the Preferred Architecture In the event that you have multiple site resilient datacenter pairs in your environment, you will need to decide if you want to have a single worldwide namespace, or if you want to control the traffic to each specific datacenter by using regional namespaces. Ultimately your decision depends on your network topology and the associated cost with using an unbound model; for example, if you have datacenters located in North America and Europe, the network link between these regions might not only be costly, but it might also have high latency, which can introduce user pain and operational issues. In that case, it makes sense to deploy a bound model with a separate namespace for each region. However, options like geographical DNS offer you the ability to deploy a single unified namespace, even when you have costly network links; geo-DNS allows you to have your users directed to the closest datacenter based on their client's IP address. pafig2 Figure 2: Geo-distributed Unbound Namespace

Site Resilient Datacenter Pair Design

To achieve a highly available *and* site resilient architecture, you must have two or more datacenters that are well-connected (ideally, you want a low round-trip network latency, otherwise replication and the client experience are adversely affected). In addition, the datacenters should be connected via redundant network paths supplied by different operating carriers. While we support stretching an Active Directory site across multiple datacenters, for the PA we recommend that each datacenter be its own Active Directory site. There are two reasons:

1. Transport site resilience via Shadow Redundancy and Safety Net can only be achieved when the DAG has members located in more than one Active Directory site.

2. Active Directory has published guidance that states that subnets should be placed in different Active Directory sites when the round trip latency is greater than 10ms between the subnets.

Server Design

In the PA, all servers are physical servers. Physical hardware is deployed rather than virtualized hardware for two reasons:

1. The servers are scaled to use 80% of resources during the worst-failure mode.
2. Virtualization adds an additional layer of management and complexity, which introduces additional recovery modes that do not add value, particularly since Exchange provides that functionality.

Commodity server platforms are used in the PA. Commodity platforms are and include:

- 2U, dual socket servers (20-24 cores)
- up to 192GB of memory
- a battery-backed write cache controller
- 12 or more large form factor drive bays within the server chassis

Additional drive bays can be deployed per-server depending on the number of mailboxes, mailbox size, and the server's scalability. Each server houses a single RAID1 disk pair for the operating system, Exchange binaries, protocol/client logs, and transport database. The rest of the storage is configured as JBOD, using large capacity 7.2K RPM serially attached SCSI (SAS) disks (while SATA disks are also available, the SAS equivalent provides better IO and a lower annualized failure rate). Each disk that houses an Exchange database is formatted with ReFS (with the integrity feature disabled) and the DAG is configured such that AutoReseed formats the disks with ReFS:

```
Set-DatabaseAvailabilityGroup <DAG> -FileSystem ReFS
```

BitLocker is used to encrypt each disk, thereby providing data encryption at rest and mitigating concerns around data theft or disk replacement. For more information, see Enabling BitLocker on Exchange Servers. To ensure that the capacity and IO of each disk is used as efficiently as possible, four database copies are deployed per-disk. The normal run-time copy layout ensures that there is no more than a single active copy per disk. At least one disk in the disk pool is reserved as a hot spare. AutoReseed is enabled and quickly restores database redundancy after a disk failure by activating the hot spare and initiating database copy reseeds.

Database Availability Group Design

Within each site resilient datacenter pair you will have one or more DAGs.

DAG Configuration

As with the namespace model, each DAG within the site resilient datacenter pair operates in an unbound model with active copies distributed equally across all servers in the DAG. This model:

1. Ensures that each DAG member's full stack of services (client connectivity, replication pipeline, transport, etc.) is being validated during normal operations.
2. Distributes the load across as many servers as possible during a failure scenario, thereby only incrementally increasing resource use across the remaining members within the DAG.

Each datacenter is symmetrical, with an equal number of DAG members in each datacenter. This means that each DAG has an even number of servers and uses a witness server for quorum maintenance. The DAG is the fundamental building block in Exchange 2016. With respect to DAG size, a larger DAG provides more redundancy and resources. Within the PA, the goal is to deploy larger DAGs (typically starting out with an eight member DAG and increasing the number of servers as required to meet your requirements). You should only create new DAGs when scalability introduces concerns over the existing database copy layout.

DAG Network Design

The PA leverages a single, non-teamed network interface for both client connectivity and data replication. A single network interface is all that is needed because ultimately our goal is to achieve a standard recovery model regardless of the failure - whether a server failure occurs or a network failure occurs, the result is the same: a database copy is activated on another server within the DAG. This architectural change simplifies the network stack, and obviates the need to manually eliminate heartbeat cross-talk.

Note: While your environment may not use IPv6, IPv6 remains enabled per IPv6 support in Exchange.

Witness Server Placement

Ultimately, the placement of the witness server determines whether the architecture can provide automatic datacenter failover capabilities or whether it will require a manual activation to enable service in the event of a site failure. If your organization has a third location with a network infrastructure that is isolated from network failures that affect the site resilient datacenter pair in which the DAG is deployed, then the recommendation is to deploy the DAG's witness server in that third location. This configuration gives the DAG the ability to automatically failover databases to the other datacenter in response to a datacenter-level failure event, regardless of which datacenter has the outage. If your organization does not have a third location, consider placing the witness in Azure; alternatively, place the witness server in one of the datacenters within the site resilient datacenter pair. If you have multiple DAGs within the site resilient datacenter pair, then place the witness server for all DAGs in the same datacenter (typically the datacenter where the majority of the users are physically located). Also, make sure the Primary Active Manager (PAM) for each DAG is also located in the same datacenter.

Data Resiliency

Data resiliency is achieved by deploying multiple database copies. In the PA, database copies are distributed across the site resilient datacenter pair, thereby ensuring that mailbox data is protected from software, hardware and even datacenter failures. Each database has four copies, with two copies in each datacenter, which means at a minimum, the PA requires four servers. Out of these four copies, three of them are configured as highly available. The fourth copy (the copy with the highest Activation Preference number) is configured as a lagged database copy. Due to the server design, each copy of a database is isolated from its other copies, thereby reducing failure domains and increasing the overall availability of the solution as discussed in DAG: Beyond the "A". The purpose of the lagged database copy is to provide a recovery mechanism for the rare event of system-wide, catastrophic logical corruption. It is not intended for individual mailbox recovery or mailbox item recovery. The lagged database copy is configured with a seven day ReplayLagTime. In addition, the Replay Lag Manager is also enabled to provide dynamic log file play down for lagged copies when availability is compromised. By using the lagged database copy in this manner, it is important to understand that the lagged database copy is not a guaranteed point-in-time backup. The lagged database copy will have an availability threshold, typically around 90%, due to periods where the disk containing a lagged copy is lost due to disk failure, the lagged copy becoming an HA copy (due to automatic play down), as well as, the periods where the lagged database copy is re-building the replay queue. To protect against accidental (or malicious) item deletion, Single Item Recovery or In-Place Hold technologies are used, and the Deleted Item Retention window is set to a value that meets or exceeds any defined item-level recovery SLA. With all of these technologies in play, traditional backups are unnecessary; as a result, the PA leverages Exchange Native Data Protection.

Office Online Server Design

At a minimum, you will want to deploy two Office Online Servers in each datacenter that hosts Exchange 2016 servers. Each Office Online Server should have 8 processor cores, 32GB of memory and at least 40GB of space dedicated for log files.

Note: The Office Online Server infrastructure does not need to be exclusive to Exchange. As such, the hardware guidance takes into account usage by SharePoint and Skype for Business. Be sure to work with any other teams using the Office Online Server infrastructure to ensure the servers are adequately sized for your specific deployment.

The Exchange servers within a particular datacenter are configured to use the local Office Online Server farm via the following cmdlet:

```
Set-MailboxServer <East MB Server> WACDiscoveryEndPoint https://oos-east.contoso.com/hosting/discovery
```

Summary

Exchange Server 2016 continues in the investments introduced in previous versions of Exchange by reducing the server role architecture complexity, aligning with the Preferred Architecture and Office 365 design principles, and improving coexistence with Exchange Server 2013. These changes simplify your Exchange deployment, without decreasing the availability or the resiliency of the deployment. And in some scenarios, when compared to previous generations, the PA increases availability and resiliency of your deployment. Ross Smith IV Principal Program Manager Office 365 Customer Experience