# Clustering

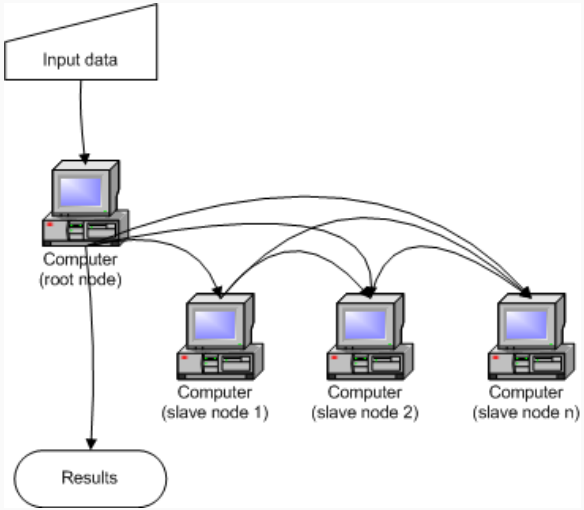
A presentation by Matt Drummond

# Outline

- What is clustering?
- Why use it?
- Where is it used?
- Implementations
- The Future

# What is Clustering?

- Independent workstations/servers working together in groups referred to as ‘nodes’ to make one cost effective, resilient computing system



# Why use Clustering?

## Fault Tolerance

- Having multiple computers allow a seamless always on experience for users
- Computationally intensive services become tolerant to unexpected stoppages

## Parallel Computing

- Allows heavy workloads to be shared across multiple computers

# The Beginning of the Server Clustering Relationship

- First implementation introduced in Windows NT Server 4.0 Enterprise Edition (1996) in the form of Microsoft Cluster Server (MSCS) that created failover capabilities
- Modifications and additions added to Windows Server 2003 with Windows Compute Cluster Server 2003 that allowed for greater performance via parallel computing (June 2006)

# Implementations

Windows Server

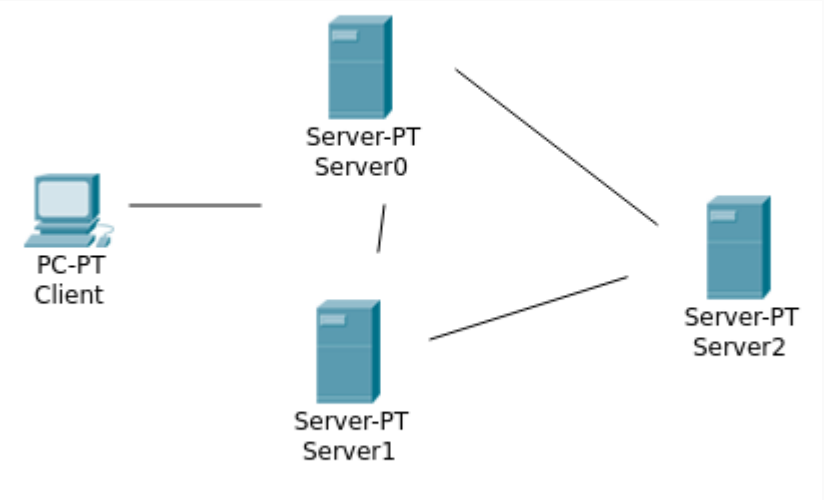
- Back End Cluster
- Front End Cluster

Alternatives

- Hadoop

# Front End Cluster

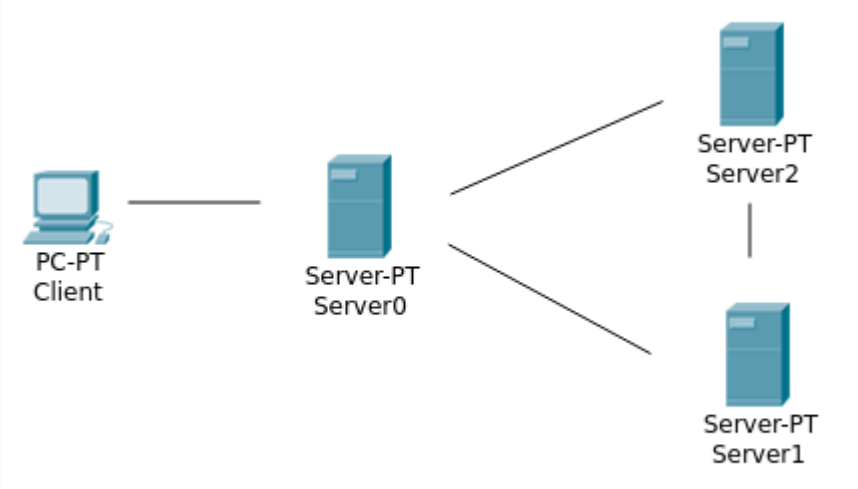
- Handles user traffic
- Focuses on network load balancing and accessing the service itself
- Front-end servers are stateless



# Back End Cluster

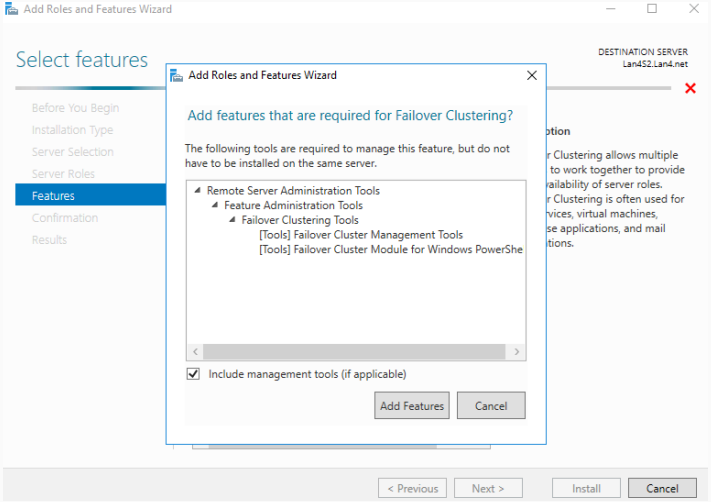
The act of having multiple nodes for the service being accessed, such as a database service

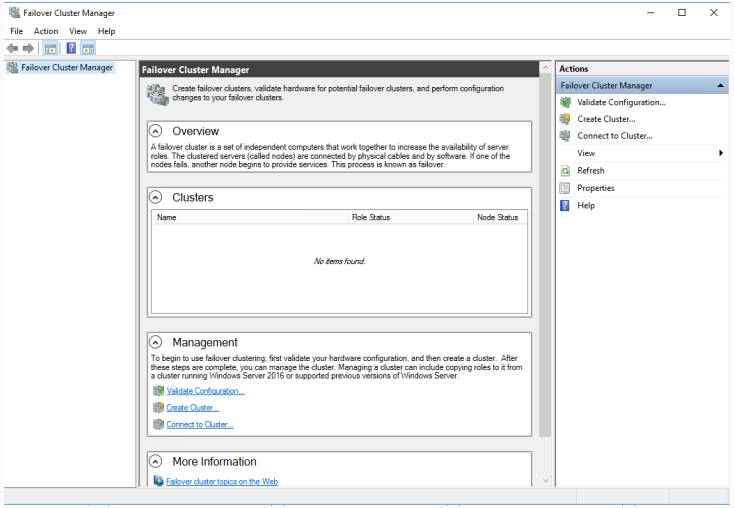
- Back end clustering provides a failover for the service itself
- Allows fault tolerance for the processes users rely on



# Failover Clustering

- Having multiple computers work together to provide fault tolerance
- Most recent implementation of clustering in Windows server
- Accessed via “Add roles and Features” in Server Manager



# Hadoop

Similar to Microsoft's idea of clustering, but with some changes:

- Open implementation of Google's MapReduce
- Maintained by Apache foundation
- Provides the backend for the top companies: Amazon, Alibaba, Google, Twitter, and universities across the world

*"We use Apache Hadoop for content generation, data aggregation, reporting, analysis"  -Spotify*

1650 node cluster : 43,000 virtualized cores, ~70TB RAM, ~65 PB storage

# The Future

- With the increasing usage of data intensive applications, fault tolerance and parallel computing will become essential in the future
- With increasing demands from users for higher network performance, clustering will become more prevalent

# In Conclusion

- Clustering can be thought of as taking multiple computers and combining them into one super computer
- Front end clustering provides load balancing to the users accessing services
- Back end clustering provides redundancy to the services themselves
- Without clustering most of today’s websites, and services would be impossible to maintain, and would have no resilience to disruptions.