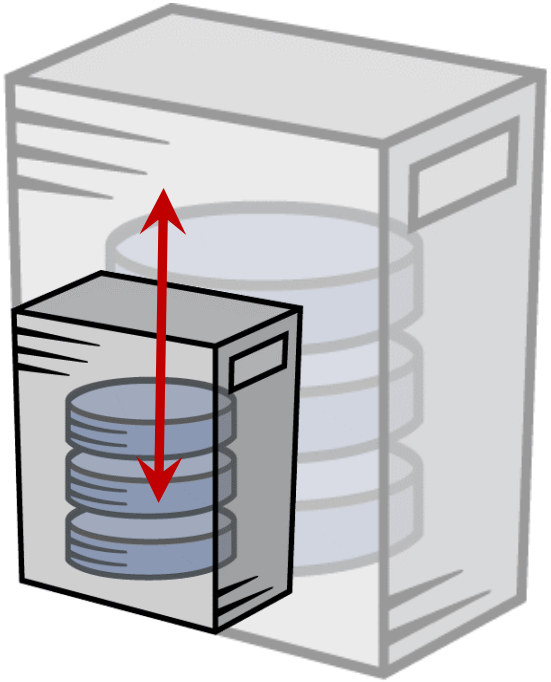
Virtualization

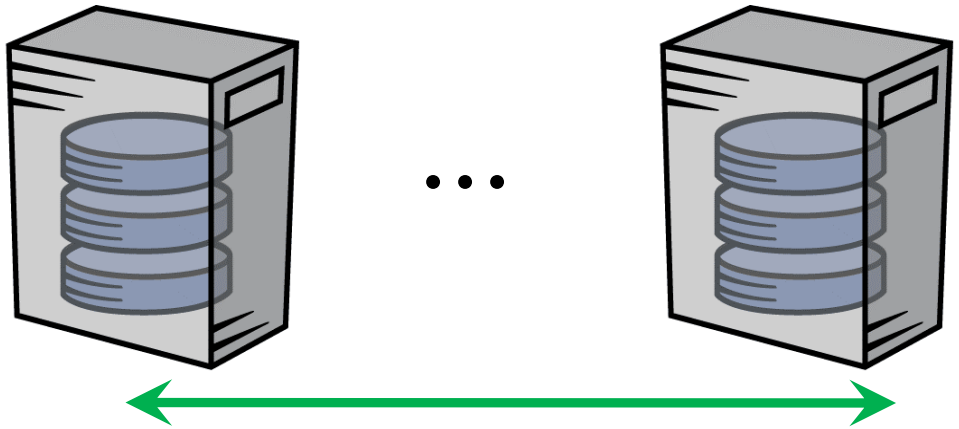
"It was much nicer before people started storing all their data in the Cloud."

Hardware Scalability

- Vertical scaling:
  expand machine



- Horizontal scaling:
  more (smaller) machines



Vertical Scaling: Supercomputer

[computerhistory.org]

Horizontal Scaling: Cluster

- Goal: more compute power, fault tolerance – cheap
  - Commodity hardware
- Approach: horizontal scalability
- cluster = (loosely or tightly) connected computers working together, appearing as single system
  - each node same task
  - clustering middleware = software controlling & scheduling
- Related
  - Amdahl's Law: predict theoretical speedup when using multiple processors
  - more recently: Playstation clusters, Xbox clusters

Horizontal Scaling: Beowulf Cluster

[Hoffman & Hargrove, ORNL]

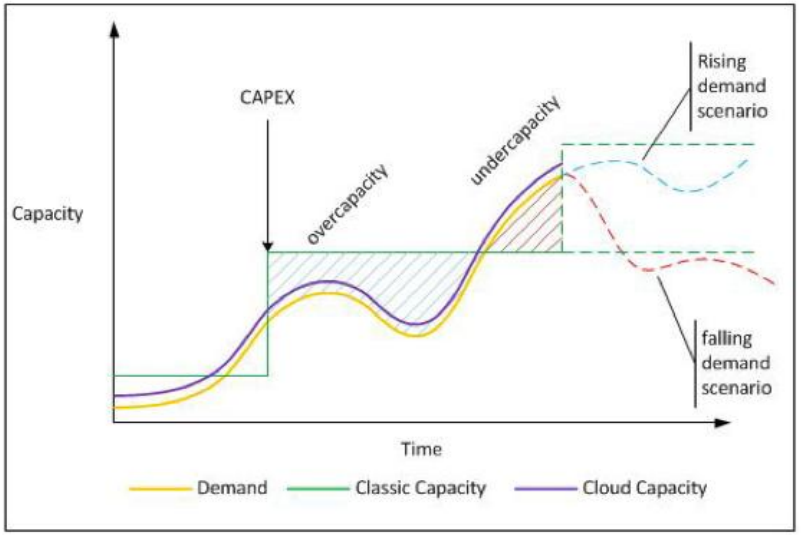
Horizontal Scaling: Supercomputers Today

TaihuLight: 10,649,600 cores in 40,960 nodes; 1,3 TB RAM; 93 PFlop/s

[top500.org / Natl Supercomputing Center, Wuxi, China]

Virtualization

- Problem: just-in-time resource provisioning
- Approach:
  - Outsourcing to service provider
  - Virtual Machine (VM) to share computer resources on demand
- Many commercial providers
  - including Amazon AWS, Microsoft Azure, T-Systems, Hetzner, ...
- IaaS, PaaS, SaaS, ...

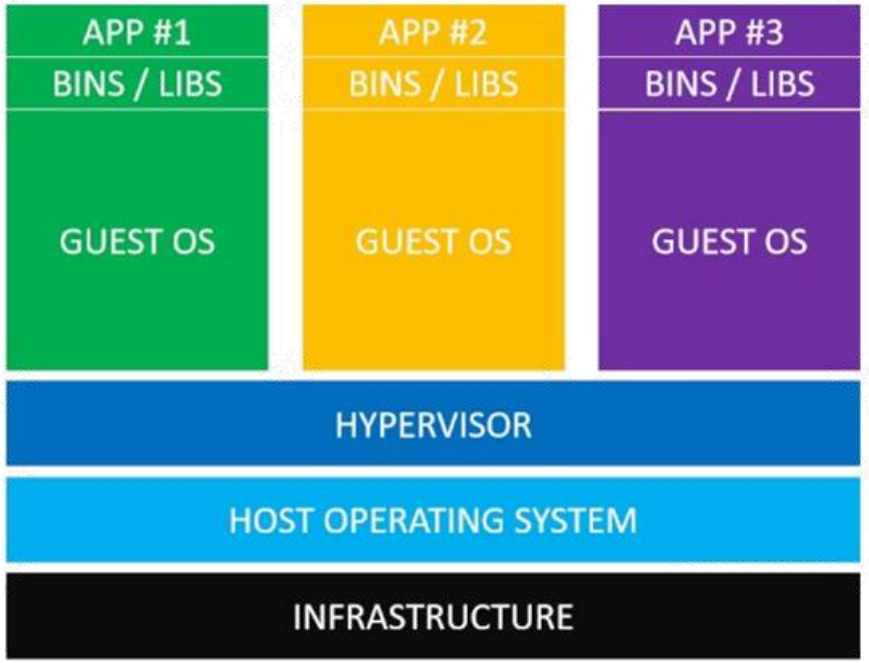


Capacity vs Utilization curves [8]

[rackspace.com]

Virtual Machines

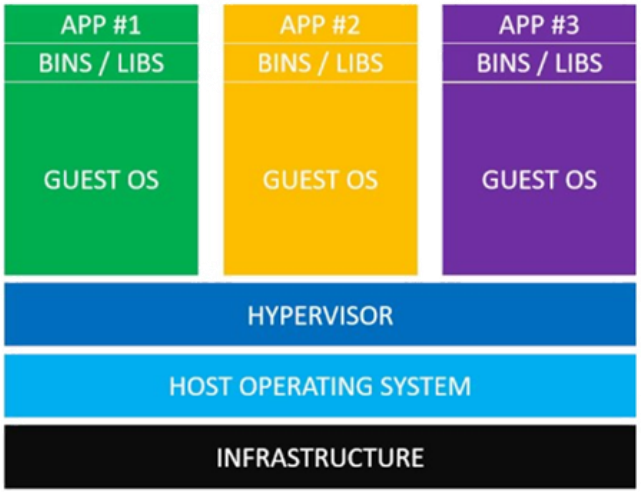
- Virtual Machine (VM) = computer application resembling a complete "computer"
  - Host system running 1..* guest systems
- Technically:
  - application invokes guest OS services
  - Guest OS calls intercepted, forwarded to host OS
  - Host OS fulfills request
- Hypervisor = virtual machine monitor
  - resource orchestration: VM start, operation, stop, ...
  - Running on host
- Data can be local or mounted from remote (ex: SAN)



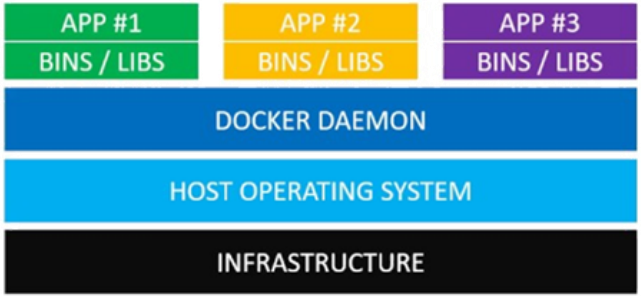
[datavizz.in]

Virtual Machines vs Containers

- Problem: Large VM overhead of Virtual Machine
  - Oversized: most parts not needed → launch time ~1min, costly updates
- Approach: Containerization = link only parts required
- PS: Microservice idea lifting off!



Virtual Machines

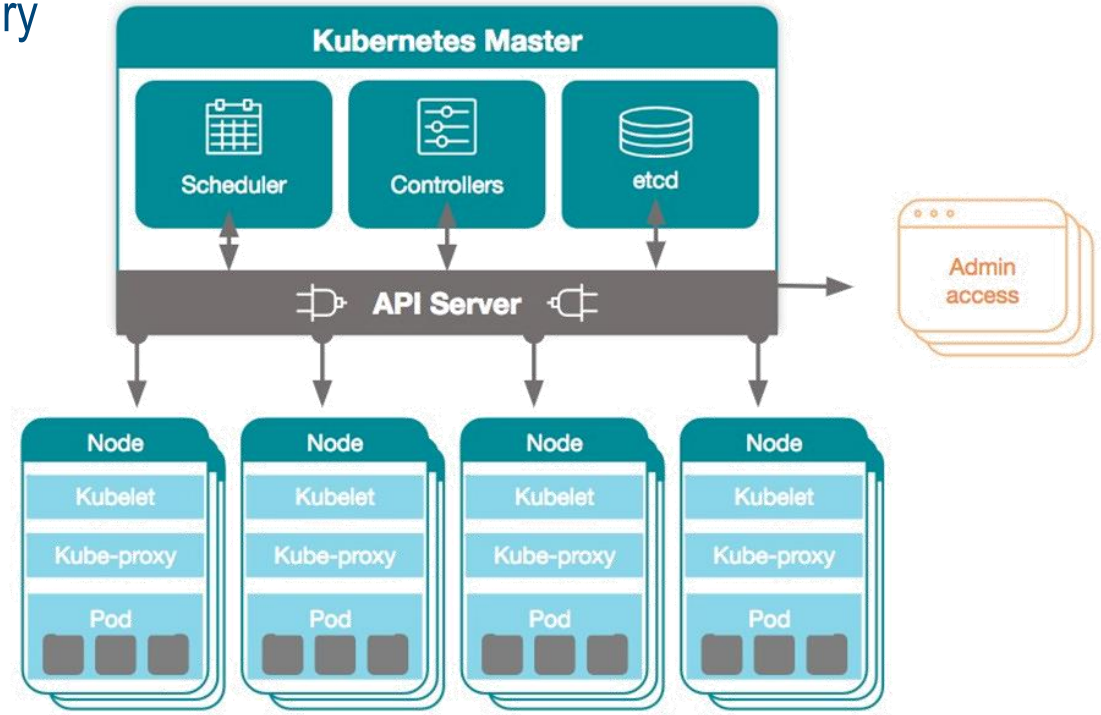


Docker Containers

[datavizz.in]

Kubernetes

- automating deployment, scaling, management of containerized applications
- group containers that make up an application into logical units
  - easy management & discovery
- Open source by Google: kubernetes.io

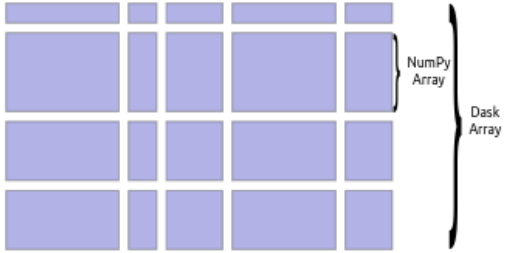


[blog.newsrelic.com]

Dask

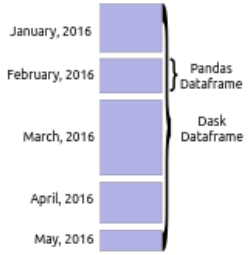
- parallelism for python analytics, enabling performance at scale
  - Dynamic task scheduling
  - "Big Data" collections larger-than-memory / distributed environments
- Open source: dask.org

Numpy



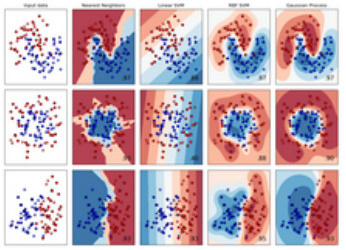
Dask arrays scale Numpy workflows, enabling multi-dimensional data analysis in earth science, satellite imagery, genomics, biomedical applications, and machine learning algorithms.

Pandas



Dask dataframes scale Pandas workflows, enabling applications in time series, business intelligence, and general data munging on big data.

Scikit-Learn



Dask-ML scales machine learning APIs like Scikit-Learn and XGBoost to enable scalable training and prediction on large models and large datasets.