Data Centre Networks

Gaurav Gautam, Preetam Patil, Parimal Parag

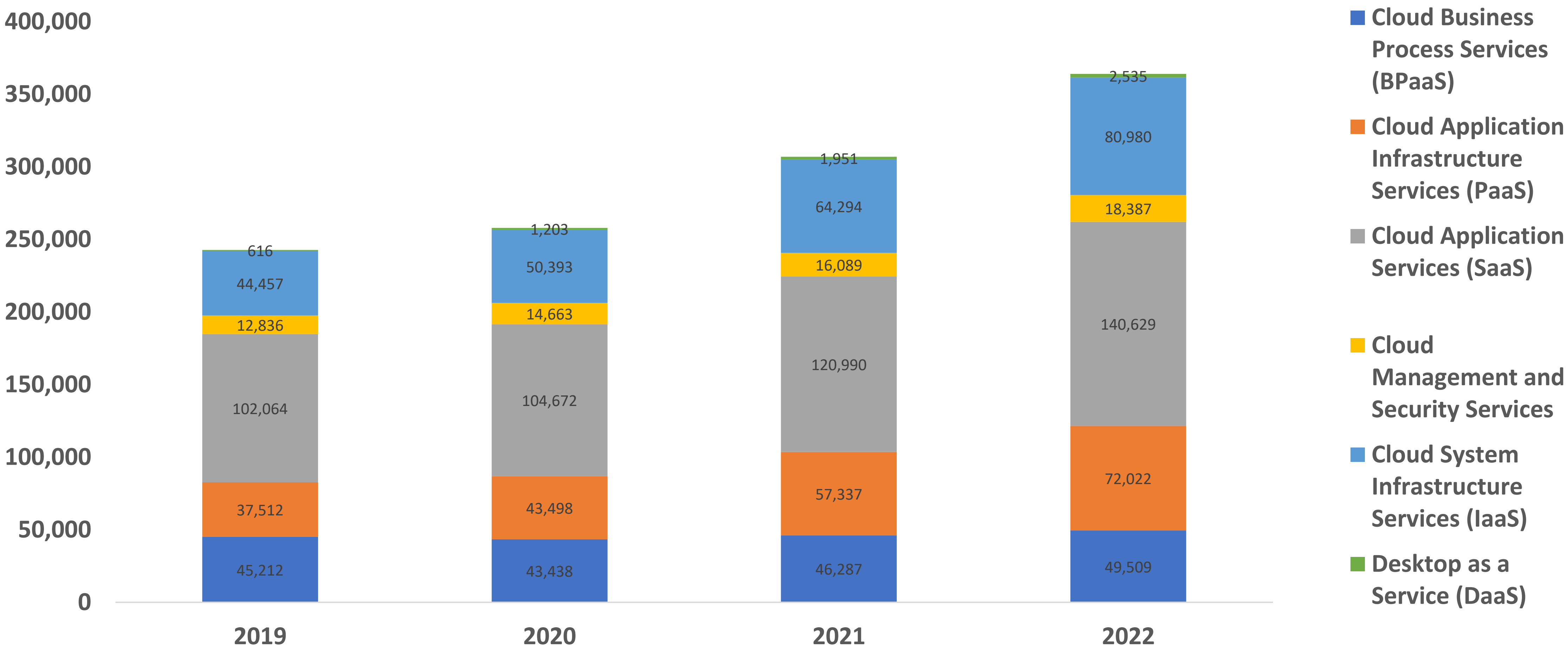
Centre for Networked Intelligence

Standing on the shoulder of the giants

Data Centre Networks

- A physical facility of networked compute and storage systems to enable the delivery of shared applications and data
- Components:
  - Network: routers, switches, firewalls
  - Storage systems
  - Compute system: servers, application-delivery controllers
- Scale:
  - Moderate scale enterprise DCN
  - Hyperscale public cloud (as of 2017, Microsoft had 1M servers in 100 data centers)

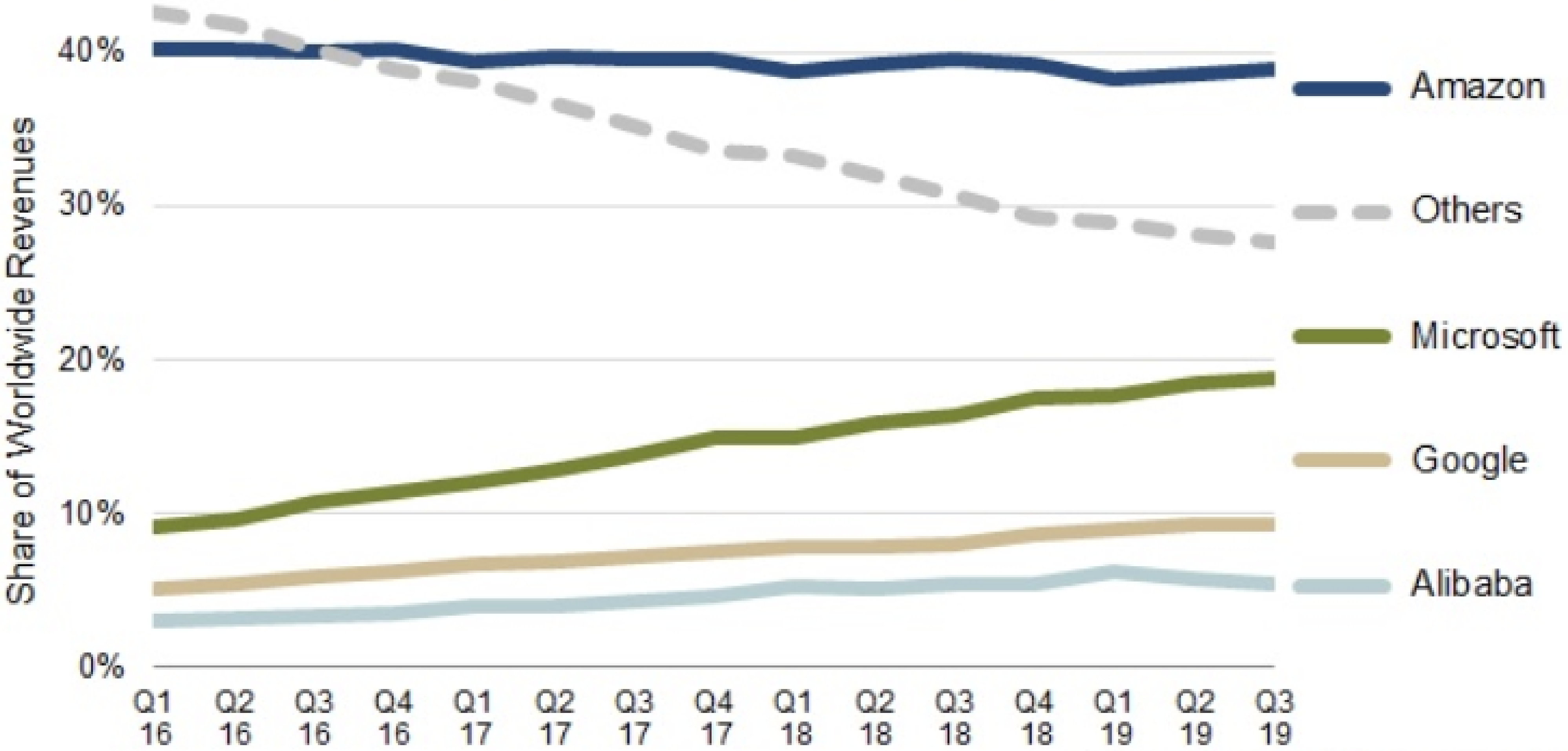
Public Cloud Revenue Growth



Source: Gartner, 23 July 2020

Public Cloud Services - Market Share Trend

(Public IaaS & PaaS - excludes Hosted/Managed Private Cloud)



Source: Synergy Research Group

DCN Design

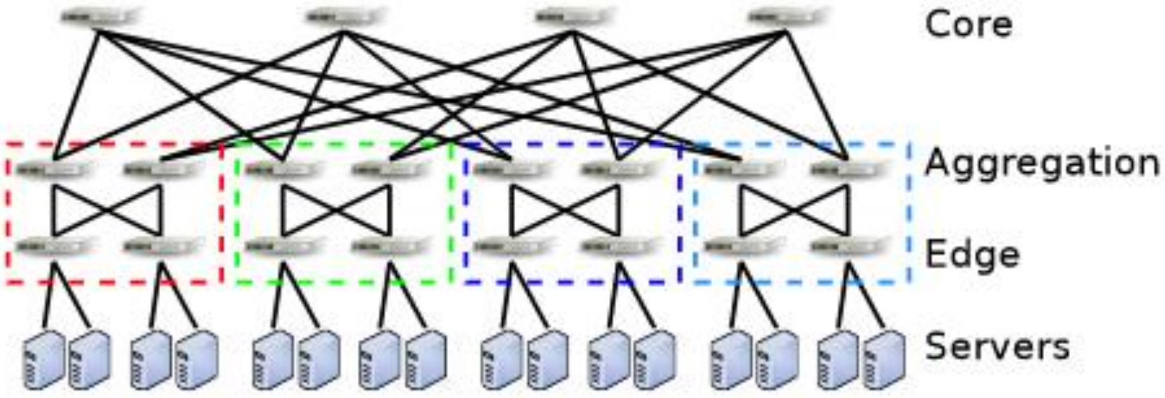
- Architecture design
  - Spine vs spineless? Multicasting? Distributed vs centralized control?
  - Modularity?
  - Programmability and flexibility vs cost?
- Data placement
  - Availability through redundancy
  - Congestion hotspot?
- Energy management
  - Cross system optimization?
- Telemetry
  - Sampling in time, network elements, features
- Traffic control

Architecture

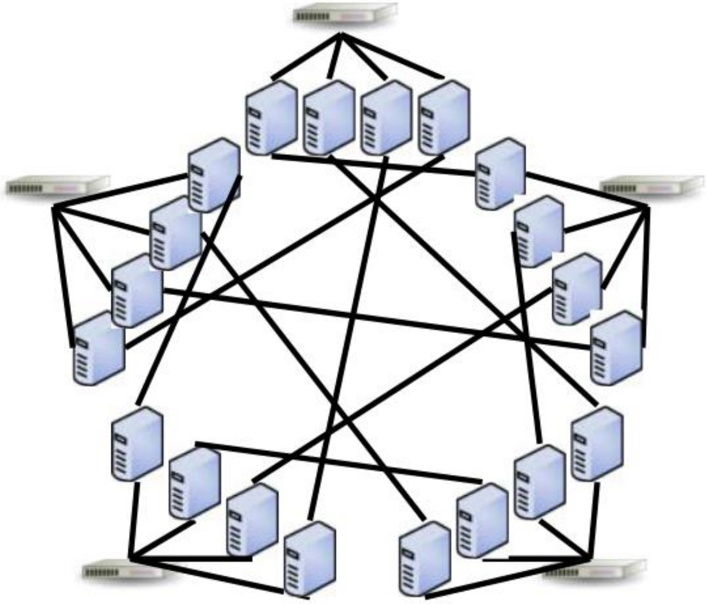
Architecture Design

- Goals:
  - Scalability
  - Availability (Redundancy in data and paths) and flexibility for fault tolerance (rerouting flows without downtime)???
  - Adaptability to load fluctuations (congestion control) and predictability (service guarantees)
  - Observability (DCN telemetry)
- Proposed architectures
  - Leaf-spine topology (better for warehouse scale DCN)
  - Spineless topology (better for moderate scale DCN)

Leaf-Spine and Spineless Topologies



4-ary Fat-Tree



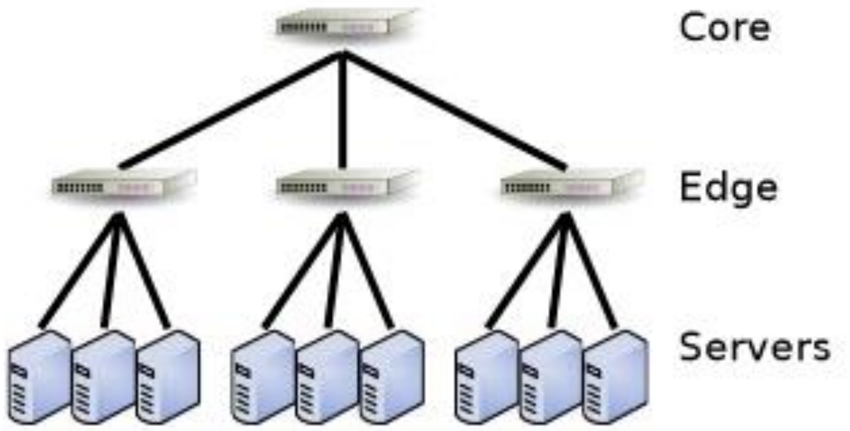
Level-1 DCell with n=4

Source: "A Survey on Data Center Networking (DCN):Infrastructure and Operations" IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 19, NO. 1, FIRST QUARTER 2017

Data Placement

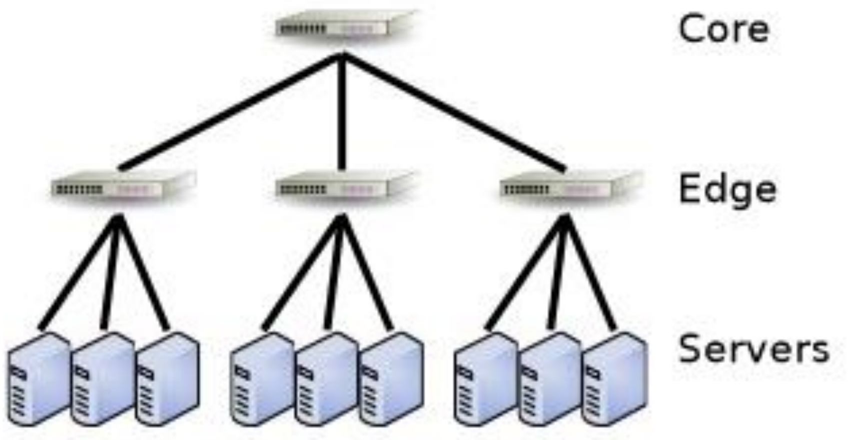
Data placement

- Goals
  - High data availability
  - High fault tolerance
  - Low storage overhead
  - Low repair and regeneration bandwidth
  - Low repair locality
  - Low access latency
- Challenges
  - Server failures
  - System overload
  - Streaming data
  - Hotspot generation
  - Large data migration

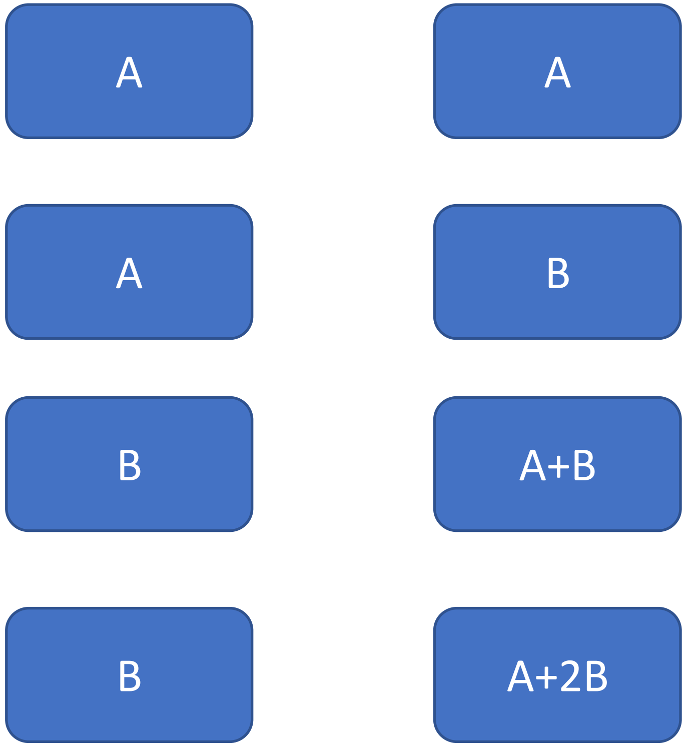


Source: "A Survey on Data Center Networking (DCN):Infrastructure and Operations" IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 19, NO. 1, FIRST QUARTER 2017

Availability through redundancy

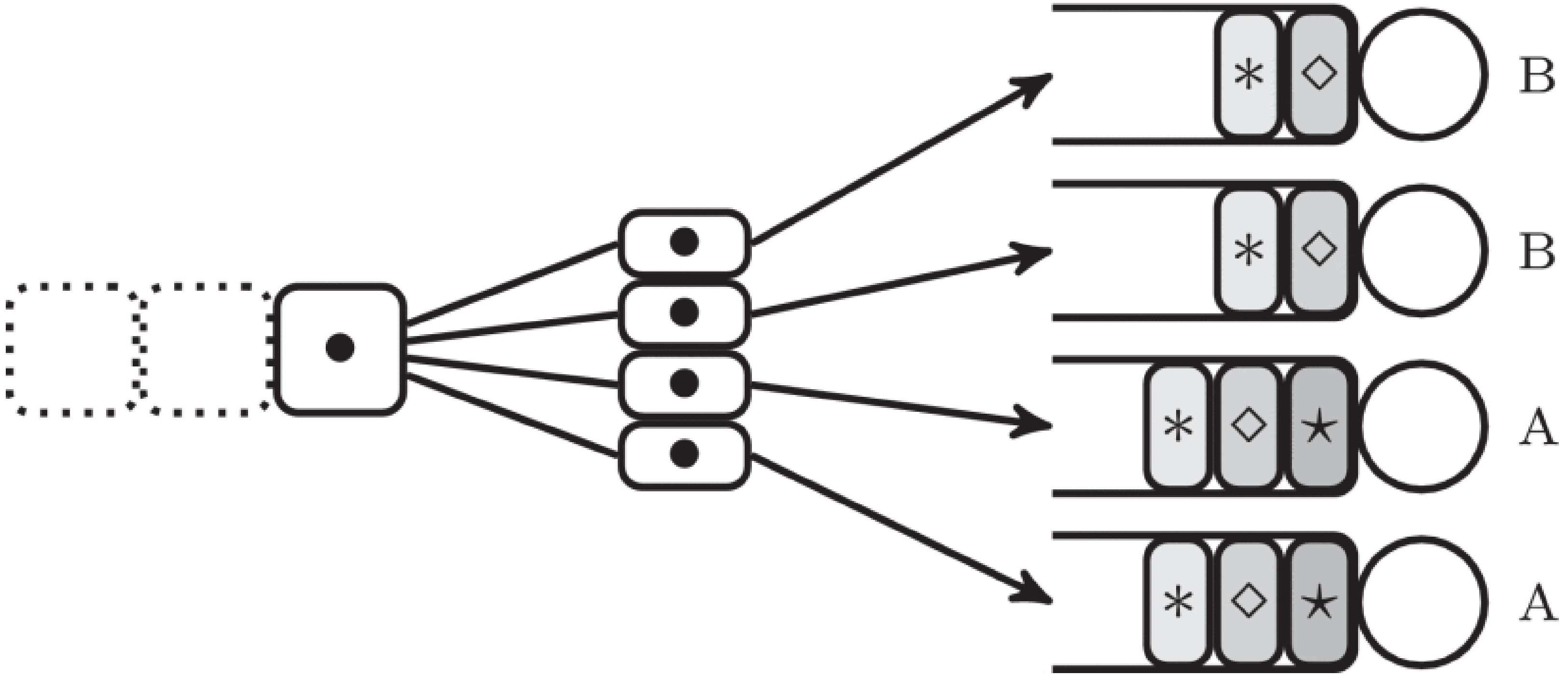


- Redundant storage of data
- Availability under finite failures
- Coding techniques for efficient storage



A	A
A	B
B	A+B
B	A+2B

Fork-Join Scheduling



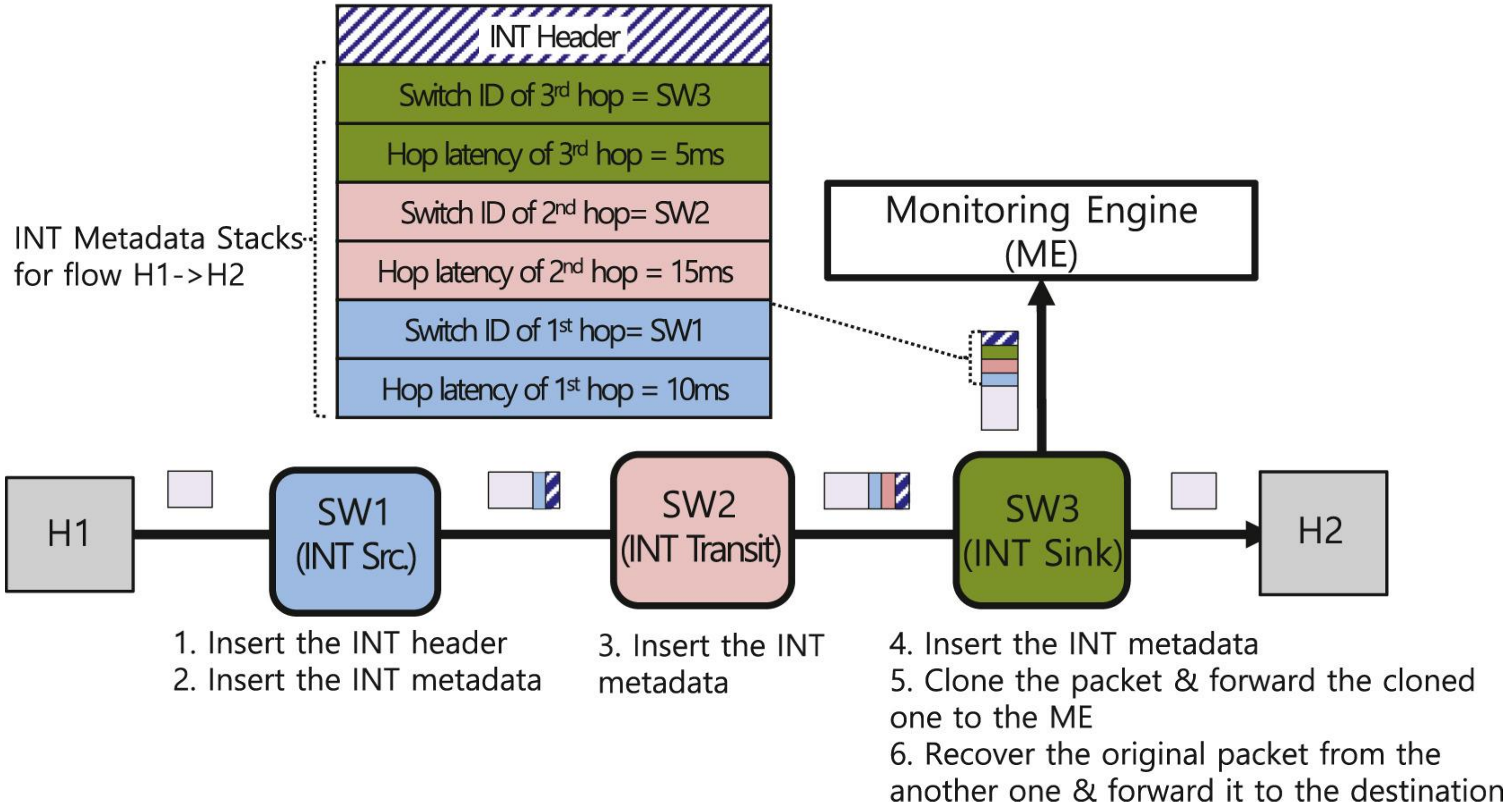
- Lower latency due to parallel access
- Higher latency due to redundant access
- Optimal redundancy selection

Telemetry

Telemetry

- Goal
  - Network state estimation with small overhead
- Measurements
  - Sampling in time
  - Sampling in network elements
  - Sampling in features
  - Raw observation vs statistics
- Typically important features
  - Latency statistics (per-flow, class, tenant for public clouds)
  - Throughput and buffer utilization
  - Energy usage vs load
- Control
  - Telemetry for control decisions such as admission/path selection

Telemetry


INT Header
Switch ID of 3rd hop = SW3
Hop latency of 3rd hop = 5ms
Switch ID of 2nd hop= SW2
Hop latency of 2nd hop = 15ms
Switch ID of 1st hop= SW1
Hop latency of 1st hop = 10ms
INT Metadata Stacks for flow H1->H2
Monitoring Engine (ME)
H1
SW1 (INT Src.)
SW2 (INT Transit)
SW3 (INT Sink)
H2
1. Insert the INT header
2. Insert the INT metadata
3. Insert the INT metadata
4. Insert the INT metadata
5. Clone the packet & forward the cloned one to the ME
6. Recover the original packet from the another one & forward it to the destination

In-band network telemetry

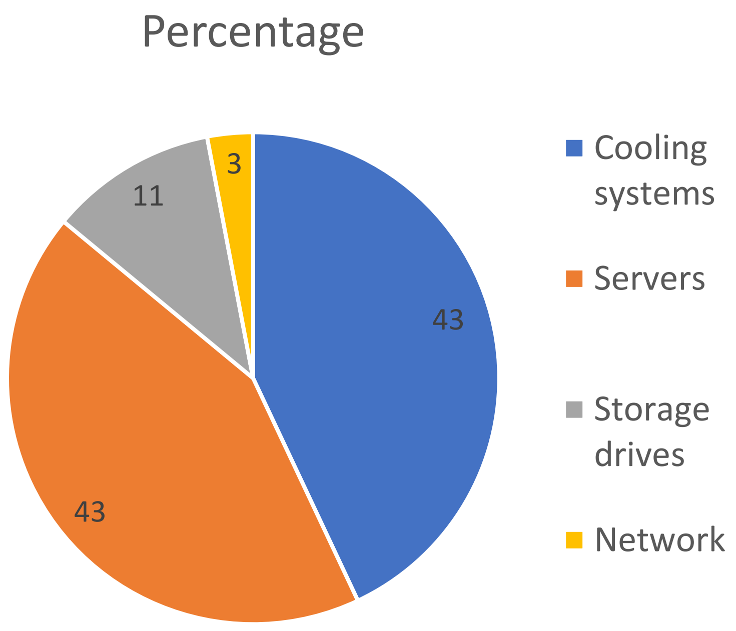
Source: "Flexible sampling-based in-band network telemetry in programmable dataplane" ICT Express Volume 6, Issue 1, March 2020, Pages 62-65

Energy Management

Source : Bloomberg

Energy Management

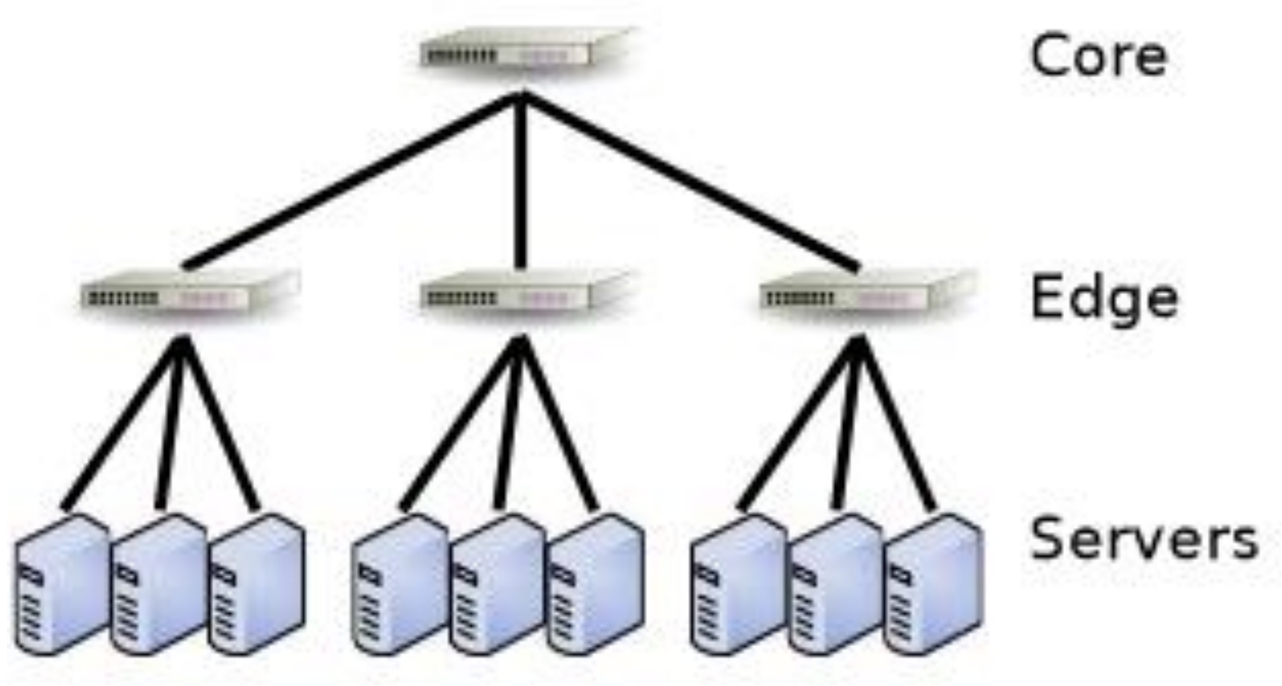
- Goal
  - Reduce the energy expenditure and carbon footprint of servers and cooling systems
- Challenges
  - Design of energy efficient storage devices
  - Design of cooling systems
- Observations
  - Hyperscale DCNs are more energy efficient



Source: Lawrence Berkeley National Lab 2016

Energy vs performance tradeoff

- Cores with several speed levels
- Faster speed higher energy consumption
- Control of server speed for energy management
- Based on network telemetry



Traffic Control

DCN Traffic Control

- Goal: Optimally utilize the available bandwidth and adapt the network dynamics
- Control decisions
    - Admission control and priority management
    - Path selection (SDN) vs routing (distributed)
    - End-to-end congestion control
- Observations
    - Complexity and overhead tradeoff with performance

Priority management

- Goal: Manage service guarantees for flows
- Flow classification
  - SLAs (tenants)
  - QoS guarantees (application requirements)
  - Elephant vs mice (best effort)
- Typical mechanisms
  - Allocate more resources by giving more scheduling opportunities
    - Priority queueing through p4
    - Priority flows go to priority buffer
  - Differential priority flows using ToS/DSCP markings in packets

Path Selection vs Routing

- Goal: Find an **optimal** path/route
- Topology discovery
  - Spanning tree algorithms
  - Multicast trees
- Centralized
  - Path selection in SDN
  - Challenges: How much network information is needed
- Decentralized:
  - Routing algorithms (BGP, OSPF, IS/IS)
  - Challenges: Sub-optimal since only local view
  - Multipath routing: SPB, TRILL (single TCP flow on a single path only)
  - Splitting a TCP flow: FLARE, MPTCP
  - Challenges: which flows to spilt, when (imbalance threshold) to split, packet reordering

End-point congestion control

- Goal: Adapt to the current network conditions to meet service guarantees
- Challenges
    - Congestion hotspots
        - TCP in-cast problem
    - Unfairness
        - Respond to higher flow completion times, throughput collapse, or Jain unfairness index
- Rate control
    - TCP flow control
        - DCTCP (utilises ECN of hardware switches)
        - ICTCP (In-cast TCP)
    - Network rate control at end-points
        - Linux: DPDK/XDP, tc, Windows: NDIS

DCN research platform implementation

- Vendor solutions:
  - DC switches (with support for segment routing, P4, MPLS)
  - Integrated analytics framework (e.g., Tetration), or custom scripts?
- Simulators: NS3, DCNs-2, DCNSim, mtCloudSim, MATLAB
- Open-source platforms:
  - ONF Trellis switching fabric with programmable switches/switching ASICs
  - Emulation: NG-SDN stack by ONF (ONOS, Stratum, bmv2)
  - Combined H/W+S/W: NG-SDN stack along-with switching ASICs and NETFPGA card (network processor), XDP-supported NICs
  - VM/Container/microVMs platforms
  - (OpenStack/OpenShift/OpenNebula/Kubernetes) for workload generation and NFV

Thanks!