# Section 1: Week 1: Distributed System Structure

Nate Bachmeier

TIM-8120: Distributed Systems

September 22nd, 2019

North Central University

# Part I: Distributed System Structure

Contoso Clothing is an international manufacturer and retailer of personal attire. Their manufacturing processes span three continents and need to meet the needs of their several thousand stores. Each location has dozens to hundreds of employees that need to have access to computing environments, point of sales services, and printers. There are multiple configuration options for the construction of this environment, each with different pros and cons. If they lack the understanding of these trade-offs, then the system will be (1) be too expensive to operate, (2) unable to meet peak loads and (3) unreliable during complex scenarios.

# Command and Control Structured

Traditional systems (see Figure 1) might use a static hierarchical structure where the headquarters will replicate policy through a tree structure (Steen & Tanenbaum, 2016). Within the tree there each branch represents an aggregation point, such as North America or Europe. There can be child branches to further distribute the load to management points, such as Washington and Spain. These management points will then execute commands on clients (leaf nodes) and collect any local results.

## Tree-Based Distribution Strategies

While these trees are good at distributing load across a vast breadth of systems, they introduce several single points of failure (Annadurai & Vijayalakshmi, 2016). To mitigate these risks, administrators implement these branch nodes as complex systems, not individual compute units (Khaneghah & Sharifi, 2013). For instance, the *Washington Management Point* might not be a single server but a load-balanced ring of servers.

By introducing a load-balancer, the administrators are trading availability for additional complexity. Consider the impact of a client sending three requests to the load-balancer, which in turn hands them to three different service instances. These scenarios can lead to (1) out-of-order eventing, (2) partial conversation failure, and (3) redundant resource allocations – to name a few challenges.

## Influence on Software Rejuvenation

However, it can simplify other scenarios such as software rejuvenation strategies as there are multiple identical processors within the functional group. Rejuvenation is the operational procedure of recycling private instance state after it has exceeded a threshold (Yang, Min, Yang, & Li, 2013). Since the functional group contains two or more nodes, the rejuvenation can be applied to one node while the other continues to service requests.

Perhaps the message processor leaks memory and becomes unresponsive after the working set exceeds 1GB. In this scenario having an external process (1) monitor the performance metric then (2) cleanly cycle the worker process as it approaches the threshold, would (3) increase the perceived reliability of the message processor.

# Transition Manufacturing Hubs to Pub/Sub

Contoso might decide that it is not economical to manage hierarchies of proprietary configuration on physical servers across their manufacturing presence. Instead, they could leverage their existing internet connections to move into a hosted cloud.

## Data Processing Networks

Data Processing Networks (DPNs) (see figure 2) is a simple and elegant pattern that leverages event publications and subscriptions to route messages (Celar, Mudnic, & Seremet, 2016) (Baudisch & Schneider, 2013). The process begins with the client posting, directly or via a broker, their notification into a multicasting service. The multicast applies any local transformations, and then forwards to zero or more subscriber’s FIFO queues. The arrival of the message triggers a *function* to process the event and remove it afterward. Generally speaking, these functions can exist in any programming language, as the transmission and implementation are decoupled. Assuming that the message within the queue represents a self-contained task description, it then becomes possible to execute each task in parallel asynchronously (Baudisch & Schneider, 2013). When combined with either Function as a Service (FaaS) or High-Performance Compute Clusters (HPCS), these parallel invocations can handle Internet-scale requirements.

## Pitfalls of Distributed System Design

Peter Deutsch formulated these flaws as the common assumptions that new distributed systems engineers make: (1) the network is reliable, secure, and homogeneous; (2) topology does not change; (3) latency is zero, bandwidth is unlimited, transport costs are free; and (4) there is only one administrator (Steen & Tanenbaum, 2016). While the Data Processing Network can address elements of these false assumptions, there are additional requirements for the systems engineers. For example, the topology is decoupled from the DPN as messages flow to topic subscribers. Similarly, the message processors are allowed to be heterogeneous as communication occurs through serializable payloads are placed in FIFO queues.

Other scenarios, such as mitigating network reliability issues, and efficiently using network capacity, require efficient protocols and procedures within the function implementation (Steen & Tanenbaum, 2016). Steen and Tanenbaum propose that these characteristics need to be locally handled instead of assuming an external third-party will “do the right thing.”

For instance, when the function pulls a message from its queue (1) there needs to be positive hand-off confirmation from (2) each of the caller services they have (3) accepted their copy of the event and (4) replicated it into a durable store. Only then, can the original message be removed from the function’s queue with confidence the payload won’t become lost. Simple scenarios, such as rejuvenation is overdue or incorrectly compensated cluster load, will cause middleware based solutions running on the caller service to become error-prone over time (Steen & Tanenbaum, 2016) (Khaneghah & Sharifi, 2013) (Yang, Min, Yang, & Li, 2013).

Contoso will also need patterns for efficiently using network resources. If customers in Asia need to wait for content located on American servers, the user experience will suffer. Or if the European customer discovers their account information has left the EU, then law-suits may follow. These classes of issues force many distributed systems to be partitioned and replicated into dedicated geographic regions.

# Retail Stores to Orchestration Services

Orchestration Services (see figure 3) are systems that execute business process workflows. They describe the sequencing requirements to perform some action, along with optional compensation actions to remediate failures (Venkatesan & Sridhar, 2015). An example implementation would be the Web Services-Business Process Execution Language (WS-BPEL) open standard running on Apache Orchestration Director Engine (ODE).

## Design Paradigm

*Programming in the Large* refers to the ability to author complex workflows using Extensible Markup Language (XML) or equivalent tags. This authoring process tends to be visual and targeted toward the business domain experts.

*Programming in the Small* refers to the code behind for these workflow tags, and any custom integration logic that is needed. This authoring process tends to target systems engineers who use their existing Integrated Development Environments (IDE).

Applications written for WS-BPEL can follow this paradigm to build highly reliable automation. The standard can also integrate into existing web service technologies such as Simple Object Access Protocol (SOAP), Representational State Transfer (REST), Web Service Description Language (WSDL), Web Service Security (WS-Security), and more (Keller, 2007). These touchpoints can significantly reduce the costs of bringing legacy systems into a more modern design.

## Challenges

Using orchestration services can simplify the development of the distributed application. However, it does not address the cluster management scenarios. Consider a workflow that manages a customer’s order flow and could run for days or weeks. In this scenario, it is not acceptable to but a throttle on the number of concurrent instances of that workflow.

When a quota does not exist, how can the cluster prevent oversubscription of resources? It cannot, or at least not directly. Instead, there needs to be a loosely coupled dependency on the executive system (Khaneghah & Sharifi, 2013). Perhaps the workflow calls the XYZ service and receives an error notification that the Service Is Too Busy. The workflow instance needs to trap that exception and perform some back off and retry policy. In parallel, the executive system needs to detect that that insufficient capacity exists and dynamically reconfigure the topology.

# Distributed Office Systems

Enterprise environments also need to support business productivity centers (see figure 4) for their employees of the company. These topologies include devices such as printers, routers, point of sales systems, and laptops.

## Configured Heirarchy

Contoso might base this environment on the Command and Control structured, with the headquarters network administration team exposing core services, such as identity, name resolution, and virtual private network (VPN) gateway. Their headquarters could then securely connect to each branch office across the VPN tunnel to deliver network policies to the branch office machines.

## Configured Security

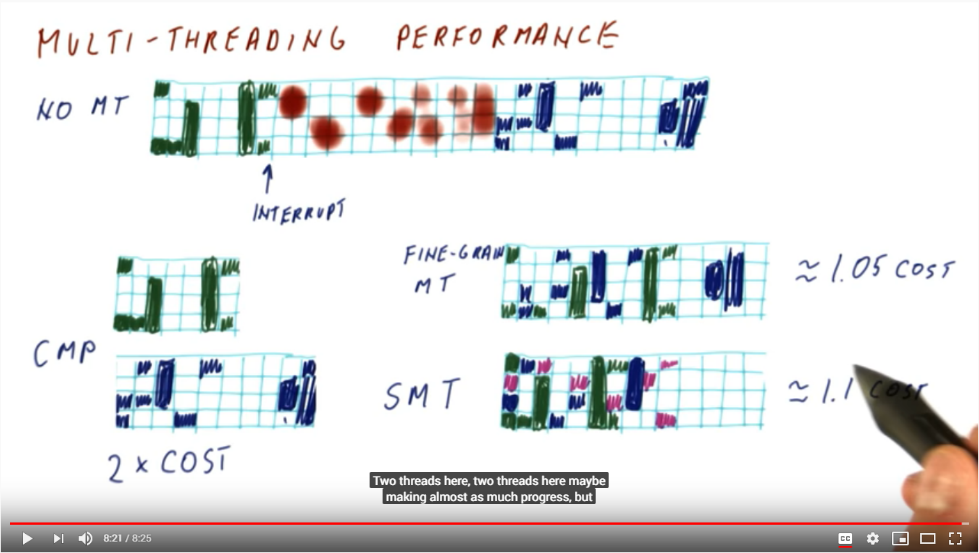
The branch office could then set a DMZ, such as a Reverse Proxy so that intranet and internet traffic remain separated. If the branch location offers Point of Sales devices, it would be advantageous to segment that portion of the network. This approach reduces the attack surface and helps to protect the resources.

# Part 2: Processor Scheduling

Modern processors can run hundreds of processes in parallel, while only containing a handful of physical cores. They accomplish this goal through efficient scheduling strategies.

## Interweaving

Georgia Tech’s High-Performance Computer Architecture course includes this diagram to explain interweaving threads (Prvulovic, 2015).



At the top is the behavior of a single core and no hardware multi-threading support. For this scenario, a thread is scheduled and runs until an interrupt. After the interrupt, context switching takes place, and for some number of cycles, no progress made. Eventually, the second thread, shown in blue, can be scheduled.

To the bottom left is the behavior for a chip multiprocessor, which uses two physical cores to run two concurrent threads. A challenge with this approach is that it is prohibitively expensive due to requiring nearly twice the silicon.

The middle right is a fine-grained multi-thread core, that attempts to schedule something in every clock cycle. These processors do not wait for interrupts and evaluate instructions as soon as they become available.

The bottom right is a simultaneous multithreaded core (SMT) and is capable of running different threads during the same clock cycle. This scenario provides the most parallel execution as work is continuously flowing to all circuits as quickly as possible.

## Pipelining

Prvulovic provides an analogy that the instruction pipeline is similar to an oil pipeline. Let’s say that you pump a gallon of oil through the pipe, and that takes one day, how long does it take to get the second gallon? The answer is near real-time, provided they started at the same point. This behavior is because both gallons took one day, but the offset between them is near zero.

The same behavior exists for instruction pipelines as the processor attempts to feed in many operations from the different threads. For instance, the Hummingbird E203 will perform the services (stage 1) finger, decode, execute, write back; (stage 2) access via LSU; and (stage 3) copy back to a general register. (Tianchuan & Zhenbo, 2019).

## Speculative Execution

Modern processors follow the SMT strategy and use instruction pipelines as described above. Their goal is to minimize the amount of idle space within that pipeline as it is time that could be used to do something productive.

One strategy is to pre-emptively execute sections of the program that are likely to follow (Kocher, et al., 2018) (Lipp, et al., 2018). If the prediction is correct, then the results for that code block are already evaluated. Otherwise, the speculative execution is not committed, and the right instructions evaluated. Since the worst case of speculative execution equals the time of having been idle, there is no harm in doing it.

With the Spectre and Meltdown vulnerabilities, it has shown that speculative execution can be used to bypass memory isolation protections and allow arbitrary memory reads. According to Kocher et al. and Lipp et al., these attacks rely on causing the pre-emptive execution to leak state into the processor cache. Though time measurement strategies, it becomes possible to determine the values leaked.

## Reorder Buffer

The densest and most complex component of the scheduling component is the reorder buffer (Choi, Park, & Jeong, 2013). This component is responsible for reassembling the out of order execution into the program defined order. Increasing the size of this structure can improve performance by increasing the amount of speculative execution; however, it also decreases power efficiency.

Choi et al. propose that one strategy is to move the reorder buffer into a separate component. Then exploit the fact that most basic blocks are around six instructions and by default correctly ordered. By optimizing for the typical case, and treating the exception as an exception, they can drastically reduce the power usage.

# Conclusion

There are multiple viable approaches to building distributed systems, and their pros and cons weighed.

Traditional methods have relied on command and structures such as tree-based distribution models. While these have been effective for decades, there are more modern solutions available through HPCS (and cloud) services. Two common patterns for building within HPCS are the Data Processing Networks and Orchestration Services approach. Data Processing Networks relies on a collection of FIFO queues that will trigger a function to perform an idempotent action. These are simple to understand and can reach enormous scales of parallel processing. Orchestration Services segment the authoring problem into the business expert and system engineer domains. Through a clear separation of duties, the right experts can be more involved with the process and not hope the other teams did the right thing. These capabilities allow defects to be surfaced sooner and reduce the impact on customers.

Just as it is not possible to have a one-size fits all distributed system architecture, a distributed system architecture could use multiple patterns. Perhaps the Retail Services portal has a more natural alignment with Orchestration and the Manufacturing Services process better aligns with Data Processing Networks. In these scenarios, it should be perfectly acceptable to choose the right tool for the job. Ultimately the goal of a distributed system to resolve a program as quickly and efficiently as imaginable (Khaneghah & Sharifi, 2013). Provided that happens and the customer is delighted, not a lot else matters.

# Figures

## Figure 1: Command and Control

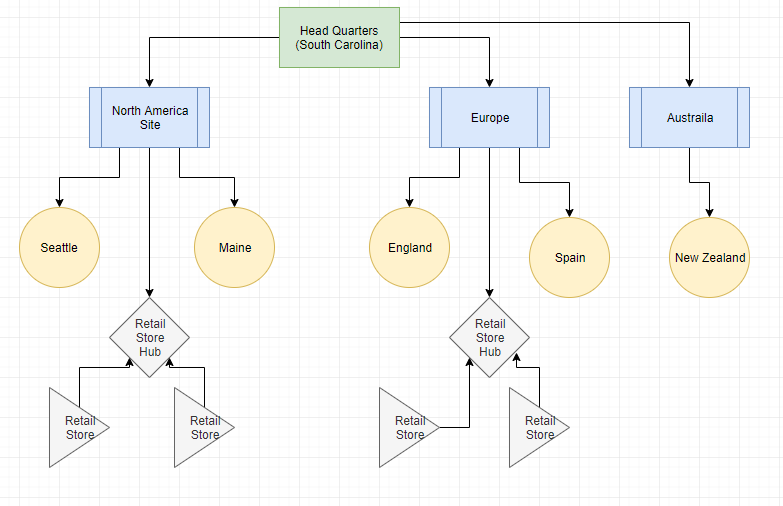
Traditional distributed systems might use a tree-based distribution model to propagate policy and aggregate results back to the top.

Figure : Command and Control

## Figure 2: Data Processing Networks

A command pattern for HPCS is to use publisher/subscription models to fan-out self-contained task descriptions. These descriptions are parallel processable, enabling enormous scalability.

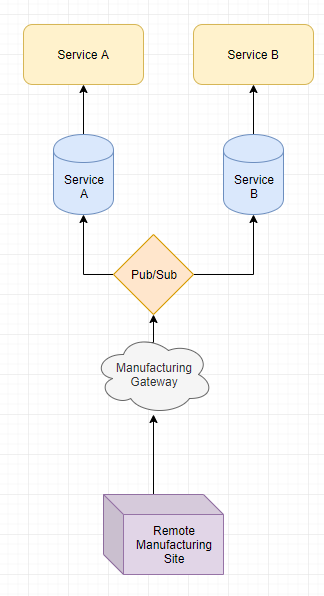


Figure : Pub/Sub

## Figure 3: Orchestration Services

This pattern leverages a central service to schedule remote calls to external services and apply compensation strategies for any failures.

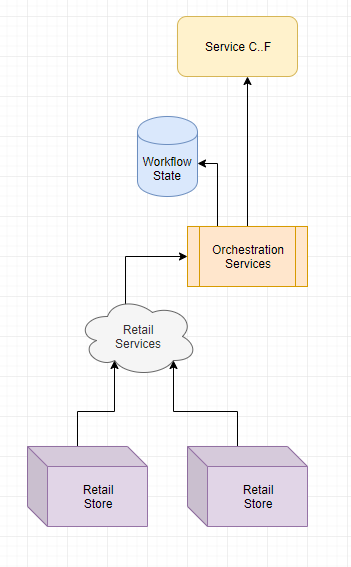


Figure : Orchestration Based

## Figure 4: Business Productivity Topology

Represents a typical corporate environment with branch office

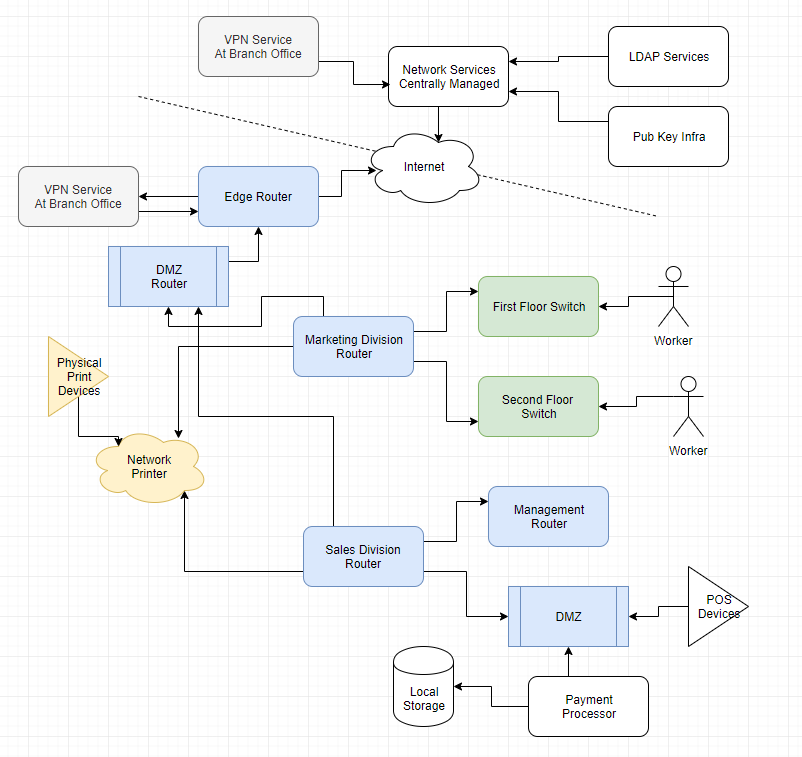


Figure Business Productivity

# References

Annadurai & Vijayalakshmi. (2016). Context Centric Cluster Computing (C4) in Ad Hoc Network.

Baudisch & Schneider. (2013). Evaluation of Speculation in Out-of-Order Execution.

Celar, Mudnic, & Seremet. (2016). State of the Art Messaging for Distributed Computing Systems.

Choi, Park, & Jeong. (2013). Revisiting reorder buffer architecture for next generation high performance computing.

Keller. (2007). *BPEL in the Real World.* Retrieved from Oasis Open.

Khaneghah & Sharifi. (2013). AMRC: an algebraic model for reconfiguration of high performance cluster computing systems at runtime.

Kocher, Horn, Fogh, Genkin, Gruss, Haas, . . . Yarom. (2018). Spectre Attacks: Exploiting Speculative Execution.

Lipp, Schwarz, Gruss, Prescher, Hass, Fogh, . . . Hamburg. (2018). Meltdown: Reading Kernel Memory from User Space.

Prvulovic. (2015). *High-Performance Computer Architecture.* Retrieved from Udacity: https://classroom.udacity.com/courses/ud007/lessons/3650589023/concepts/9999288670923

Steen, V., & Tanenbaum. (2016). A brief introduction to distributed systems.

Tianchuan, & Zhenbo. (2019). An ultra-low power RISC-V processor pipeline structure.

Venkatesan, & Sridhar. (2015). A novel programming framework for architecting next generation enterprise scale information systems.

Yang, Min, Yang, & Li. (2013). Software rejuvenation in cluster computing systems with dependency between nodes.