Scaling Integration for Modern Workloads in the Era of AI

tray.ai

CONTENTS

Modern integration and automation workloads are subject to more significant demand spikes, unpredictability, and volume than ever

The adoption of AI technologies significantly contributes to these demand spikes. As businesses integrate AI into their operations, the computational demands escalate, requiring more robust and adaptable integration solutions.

Technologists often face a choice: Plan ahead by paying for and provisioning Containers or Workers such as vCores that *should* be ready to crunch peak volume or risk processing failures and hours hunting through logs to diagnose scalability bottlenecks.

Integration scalability demands have changed

Event-driven integration is overtaking traditional polling or daily scheduled workflows. The ubiquity of Web-hook support or native triggers included in most modern cloud-based applications means collectively, a single Sales, Marketing, or Services applications stack can easily consist of 40+ apps[1] . These, in turn, can throw tens or hundreds of millions of events per second that can place massive demand on an integration platform to queue, route, transform, and aggregate data and trigger multiple downstream workflows. And as more and more enterprises adopt AI-driven processes, demand will grow exponentially.

A large enterprise can run over 1,300+ apps[2] across the organization with the number of apps increasing by 5-10% annually. Furthermore, we can expect many of these apps to incorporate AI into their platform over the next few years.

The sheer number and growth of business apps and the need for increased business process connectedness can place a massive strain on integration and automation tools, and the ops teams that support them.

Data is increasing in volume and unpredictability

Predictably, data volumes continue to increase, growing at around 25% annually, with enterprise data growing at 2X the rate of consumer data[3]. However, what's changed more is Peak Volume. It can now reach 100X of average volume for some organizations, especially when dealing with seasonal E-Commerce transactions, social spikes, IoT surges, or usage crunches from mobile internet applications. This surge is further accelerated by AI adoption, where vast datasets are required to train and operate AI models, leading to unprecedented levels of data volume and complexity.

Spiraling AI demand elevates pressure on integration scalability

AI requires data volume elasticity and scalability to effectively handle the increasing demands and complexity of data. Here's why:

Training and model development. As the complexity of AI models increases, organizations need the ability to scale up their data volume to provide sufficient training data. Elasticity allows organizations to handle large-scale data ingestion, storage, and processing required for training AI models without operational overhead.

Incorporate AI into real-time processes. AI systems often operate in real-time or near real-time scenarios where data is generated at a rapid pace. For example, in applications like fraud detection or recommendation systems, new data arrives continuously, requiring immediate processing. Integration platforms and API integrations are the vital glue but must scale to high data velocity and efficiently process it for AI-driven workflows.

Dealing with diverse and growing data. AI algorithms are designed to process and analyze diverse data types, including text, images, audio, video, and sensor data. The volume and variety of data can vary significantly, requiring elasticity to handle the dynamic requirements, not to mention data types like images and video can have significantly larger footprints than typical data.

Increasing data sources. With the proliferation of apps, databases, and even data from devices, organizations now have access to data from numerous sources. Scalability allows organizations to ingest data from various sources and handle the growing demands of data integration, so that AI algorithms benefit can deliver more accurate insights.

[1] State of Automation Survey, Tray.ai

[2] McAfee Cloud Adoption Report

[3] IDC Global DataSphere Forecast

Legacy integration platforms haven't kept pace

The expansion in apps that must be connected, increasingly real-time and data volume demands, and the growing gulf between average and peak processing means that for technologists and integration specialists, sizing and provisioning traditional integration and platforms is not sustainable. They are too complex to size, too costly to configure, requiring over-purchasing Atoms, vCores, or other worker nodes from the platform vendor. Nor to mention considerations around load balancing, parallel processing, and other areas. Additionally, the rise of AI in business apps necessitates a more modern approach to integration platforms, one that can dynamically scale to meet the high computational demands of AI processes.

All of this to meet anticipated demand which typically ends up being hard to maintain, requiring constant monitoring for failed executions, API retries, or errors related to under-sizing.

It may have been viable in an era for dozens of integrations in the enterprise, but not for the connected enterprise with hundreds or thousands of integrations at play.

Beyond IT, business teams and citizen integrators looking to connect their stack using departmental and point-to-point tools can quickly get overwhelmed as low-end tools. They can rapidly get stretched beyond what they were designed for, triggering Flood Protection and other Timeouts, creating roadblocks.

Time to rethink integration platform architecture

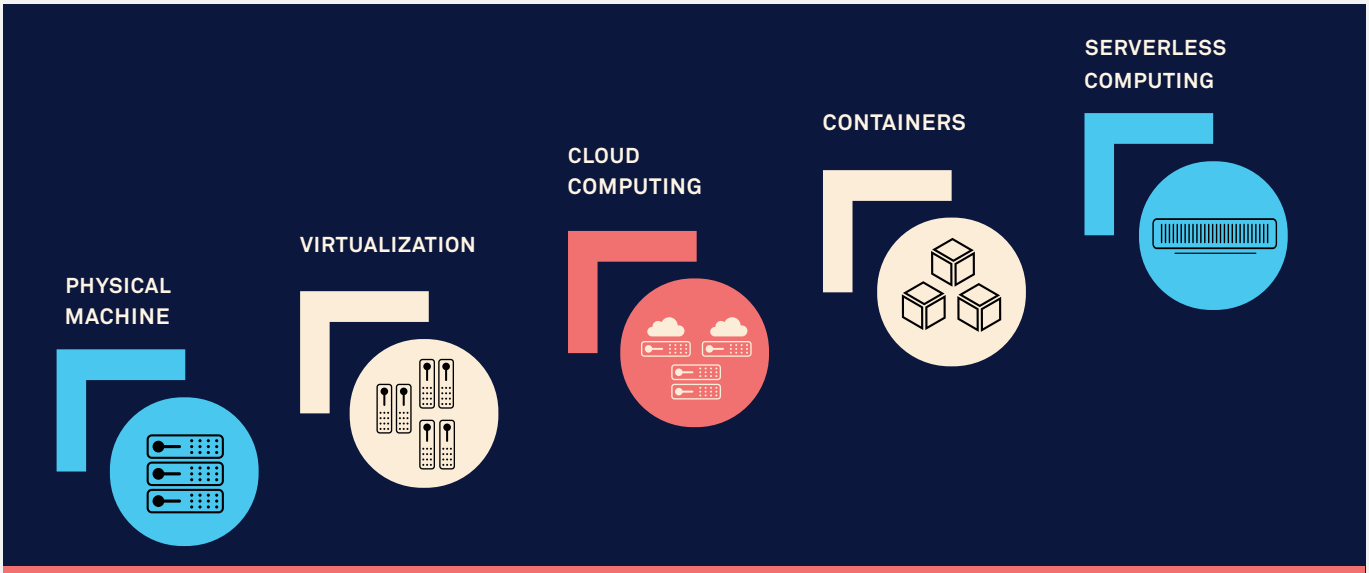
Surprisingly, some integration platforms' codebases date back nearly twenty years. As a result, they were never designed for contemporary integration demands—that require event-driven integration at scale and flexible processing to meet unpredictable volumes and demand spikes.

Instead, modern integration platforms take advantage of the latest innovations in AI, cloud-native computing, and elastic serverless processing and achieve massively parallel scale, on-demand.

The future of integration is serverless

> “By 2025, 60% of new event-driven applications will use serverless computing due to its rapid elasticity, cost agility, and low operational overhead”
>
> —Gartner[4]

Serverless computing has experienced a dramatic rise in the last five years—providing fine-grained elasticity through the decomposition of execution into highly atomic serverless functions that run on a cloud Platform-as-Service such as AWS, Microsoft Azure, or Google Cloud Platform.



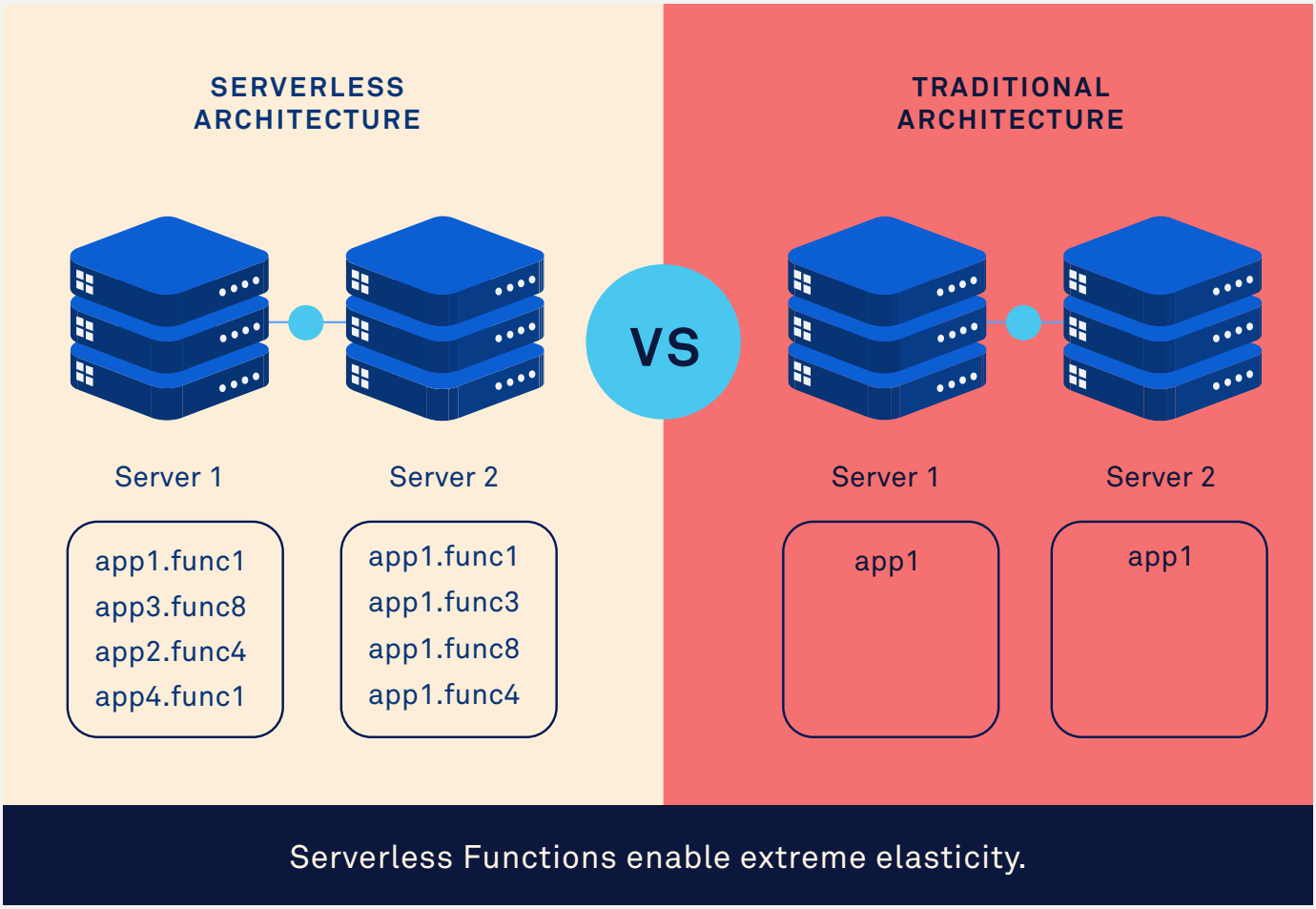
Serverless computing is the latest evolution in computing to provide more elasticity.

Serverless architectures provide the following four traits

1. The granular use of computing resources
2. Resources that do not need to be pre-allocated
3. Highly scalable and flexible
4. Users only pay for the resources they use, not purchase nodes/servers.

Serverless architectures automatically provision computing resources required to meet a workload on-demand or respond to a specific event. They automatically scale those resources up or down in response to increased or decreased demand. And then automatically scale resources to zero when the application stops running. It means the most efficient use of computing resources for customers, with the least operational overhead.

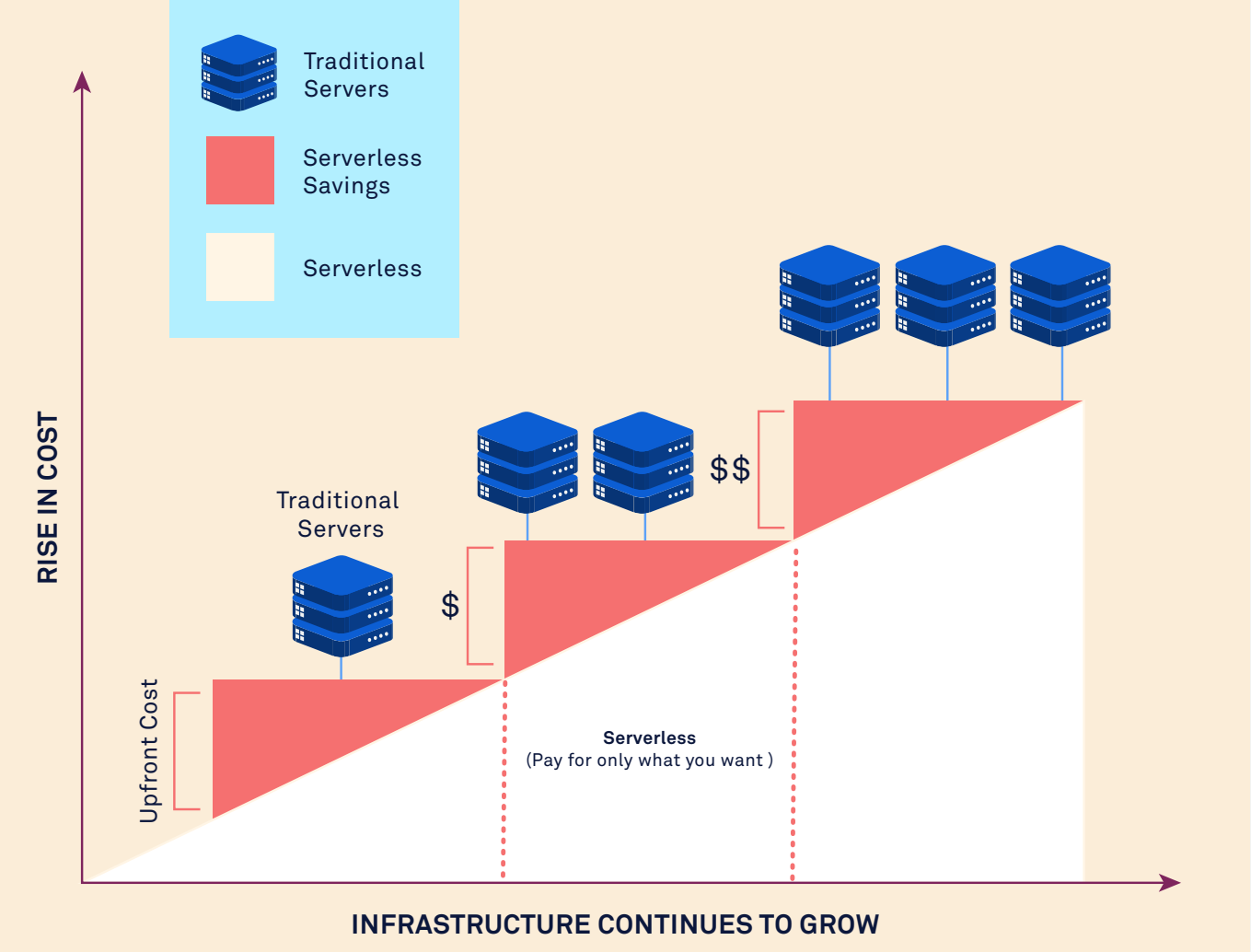
[4] **Modernizing Software Development is Key to Digital Transformation, Gartner**



Serverless computing has been adopted for various use cases, from microservices, mobile back-end processes, handling sensor data for IoT, E-Commerce apps. However, most integration platforms are pre-serverless architectures, as they were built before the invention of serverless computing.

In the context of integration, pre-serverless applications often require a higher degree of sizing expertise and pre-allocation and purchase of integration worker nodes to handle anticipated workloads. Worker nodes typically carry a per-worker cost and run a certain degree of transactional or API volume. Therefore, more must be purchased and deployed as workloads increase.

Pre-serverless architectures mean living within constraints and devoting ops resources to managing them. For example, an integration process may run smoothly for a week before it slows or crashes due to high memory usage, too much transactional volume that cannot be processed within a timeframe, or too many concurrent requests. The ops team will then examine the logs, determine if it's a resource issue, make the necessary changes, restart the worker due to a memory leak, or purchase more nodes from the integration vendor.



Serverless architectures eliminate the need to predict provisioning requirements.

While in a true serverless architecture for an integration platform, each workflow step or task in an integration workflow flow (such as a trigger, transformation, or insert) is executed as an individual Lambda serverless function and run on the underlying Platform-as-a-Service on-demand, elastically scaling in milliseconds.

There are no persistent worker nodes, and no resources are used if there is no activity. There is no need to size or pay for anticipated demand. It provides extreme elasticity.

In addition, because each workflow step is a granular serverless function executed on-demand, it provides opportunities for a high degree of concurrency with serverless workflow steps being run across a cloud platform.

	Serverless integration	Pre-serverless integration
Provisioning required	NO	YES
Cost for “Worker nodes” to scale	NO	YES
Sizing for “Worker nodes”	NO	YES
Lights out elastic processing	YES	NO

Serverless scalability case study: Eventbrite

Eventbrite (NYSE: EB) is the global self-service ticketing platform for live experiences. As a result, the company needed to deliver integrations to its customers at a massive scale. It has more than 650,000 creators on its platform, managing more than 4.6M events in nearly 180 countries. As a result, any integration Eventbrite rolls out can easily create a surge in demand on its integration platform.

9.5B
Tasks in 10 months

111K
Active integrations

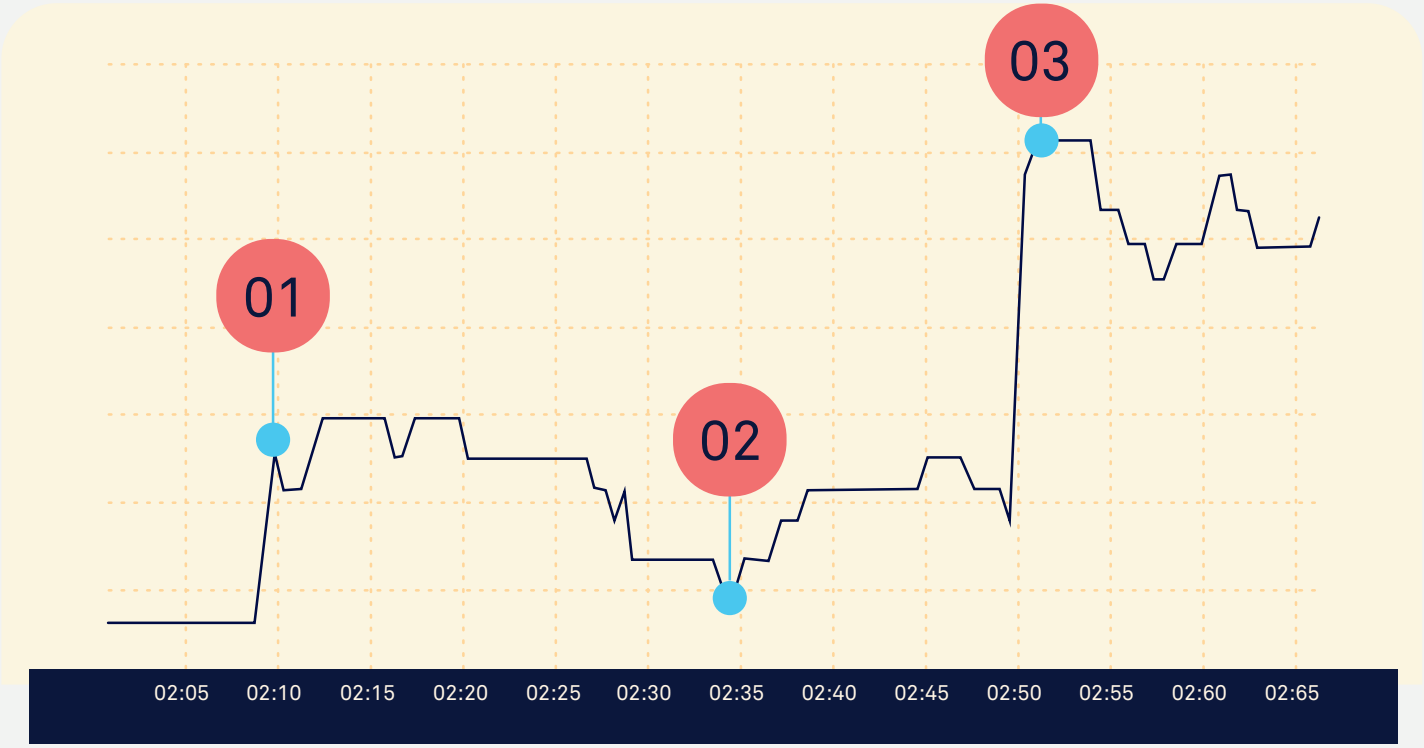
1.38M
End-customer workflows

75%
Reduction in operational overhead

By choosing the serverless Tray Universal Automation Cloud, Eventbrite enabled more than 111,000 active customer integrations in less than 12 months without adding any staff to its operations team. So, while it was able to reduce the number of engineers for each integration from six to one—more importantly, Eventbrite avoided the massive operational overhead from integrations at scale that comes with older, non-serverless integration platforms.

Serverless elastic processing in action with Tray Universal Automation Cloud™

The ideal elastic integration architecture scales dynamically and instantly without intervention or pre-allocation of resources. The real-life example below shows how this can play out on Universal Automation Cloud’s serverless architecture, lights out, over one hour.



Tray Universal Automation Cloud instantly scales with workloads without IT/Op's planning/intervention.

At (1) customer integration transactional load unexpectedly triples before 2:10 AM. The serverless Automation Cloud automatically triggers workflows that consist of serverless Lambda functions on AWS that elastically accommodate demand.

At (2) 2:35 AM, processing volume demand halves, scaling down compute, however shortly after, it triples again at (3) 2:50 AM. Again, no pre-planning or operational intervention is required.

A pre-serverless architecture would have required pre-allocating enough resources and purchasing enough workers to handle the burst of peak volume at (3). If the ops team had only planned for first peak at 2:10AM, and bought/provisioned a small bonus over that, then integrations may have failed temporarily.

The Automation Cloud scales so elastically because the unit of processing for the platform is a Task, a coarse-grained or fine-grained workflow step, a serverless function that consumes resources when needed, and zero resources when the work is complete.

Tasks can include an integration, multiple API calls, multiple transactions, a message/event, multiple rows/pages of data, a processed document. Tray.ai's large enterprise customers routinely process ~10 billion Tasks per month and have processed up to ~20 billion Tasks per Month, without requiring provisioning or sizing a priori. In some cases, their demand is highly seasonal, only spiking to billions of tasks for a fraction of a period.

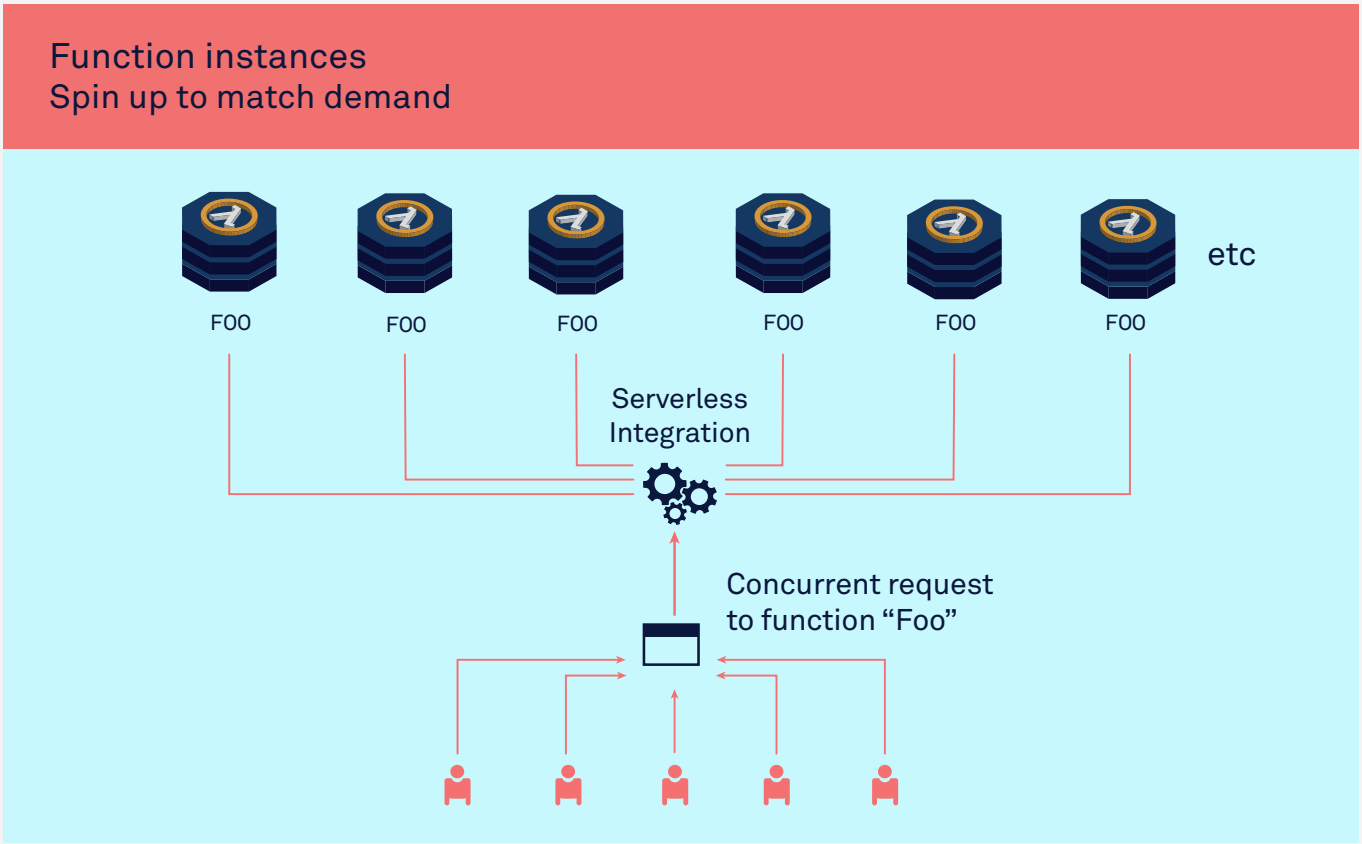
> “MuleSoft has a large overhead, and we would have required skills, a dedicated team, and a strong ongoing sustenance model. Tray Universal Automation Cloud is the other way: more user-friendly and easy to scale up. It can be managed by one or two internal resources with Tray’s support.”
>
> **—Sunita Raja, Head of Business Technology, Udemy**

The future of integration is highly parallel

The shift to event-driven integrations, where workflows can be triggered thousands of times in a matter of seconds and where large numbers of them may run concurrently for a portion of their runtime, can be demanding on inelastic architectures.

It’s where true serverless architectures shine because each workflow instance is effectively a series of Lambda functions—so whether a five or thousand simultaneous workflow instances, the compute is all on-demand.

While a pre-serverless architecture certainly provides concurrency, it is often constrained by allocating worker nodes, threading limits, compute limits, and other factors. However, keeping all of this at the ready is cost-prohibitive. And there are often constraints in older architectures, with shared resources, such as shared databases, limiting parallelism when spreading work across nodes.



Serverless integration architectures enable easier concurrency.

In contrast, a modern serverless integration architecture can dynamically spin up practically unlimited workflows consisting of Lambda functions that run for a fraction of a second. In many cases, each Lambda function is stateless, with limited contention for shared resources— providing practically unlimited parallelism.

In the case of an event-based integration, which might trigger workflow hundreds or thousands of times in a second, a serverless architecture can execute numerous workflows concurrently, massively parallel—automatically, and lights out.

Scalability case study: AdRoll shifts to real-time processing

AdRoll needed an easy, fast way to refresh hundreds of attributes for approximately 650,000 Salesforce opportunities continuously. But they were constrained to just a handful of attributes refreshing periodically using a previous custom-coded integration. This approach not only came with a high operational overhead but also led to out-of-date data in Salesforce's opportunity records.

By leveraging Tray Universal Automation Cloud's modern architecture, AdRoll now processes over five million integrations and has moved from syncing a small number of attributes to around 240 - at no increase in operations.

The future of integration is auto-scaling: Real-time + big data

As discussed, event-driven integration is on the rise, and serverless architectures are ideal for handling this kind of use case. But analytical databases like Snowflake and Redshift require bulk data loading—less high-frequency workloads, more raw transactional volume associated with Extract-Transform-Load (ETL) / Extract-Load-Transform (ELT), or Data Integration (DI) processes.

Many integration platforms are designed for one use case, triggered fairly atomic integrations (Application Integration), or bulk data integration (ETL/DI) scenarios, but not both. Organizations often use multiple tools for application and data integration due to different scalability demands.

However, the emerging demand for Reverse-ETL use cases means one platform that can scale to both app and data integration is preferable but must be flexible enough to scale with the unique characteristics of both workloads.

Reverse-ETL is where data isn't just flowing into data warehouses like Redshift or Snowflake for analytics it's also flowing out of them to drive business processes on-demand to drive activities like personalized email campaigns or website product recommendations.

Reverse ETL: Where bulk loading meets on-demand business process

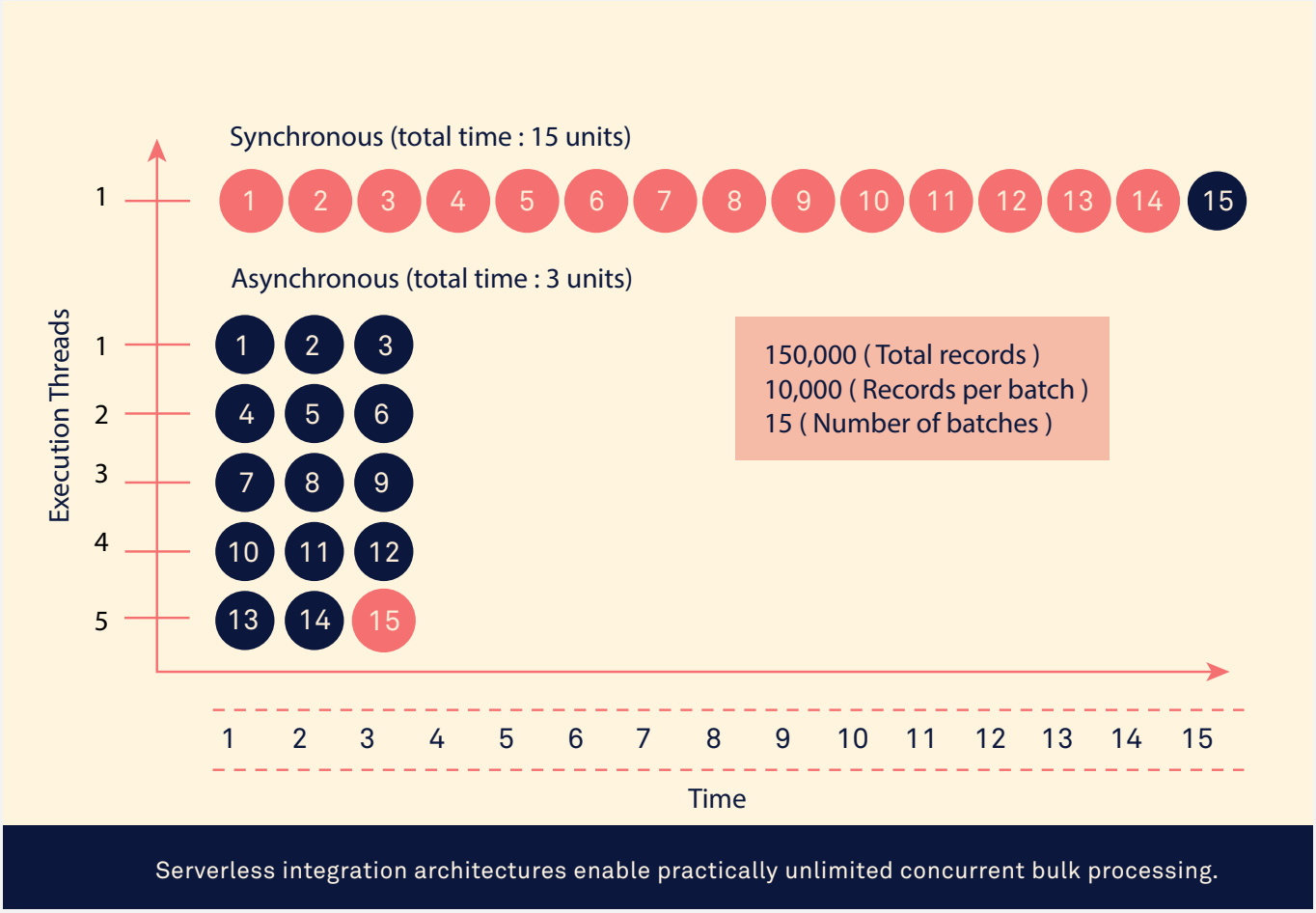
Modern serverless integration architectures enable highly parallel bulk processing

Processing large volumes of data not only means sheer volume in terms of the number of records and throughput, but it also often means complex processing (such as aggregation or enrichment) operations. This is essential for AI-centric workloads.

The rise of AI has brought about a profound increase in data volumes from ingestion, to being put to work in workflows and processes across the business. The advent of GenAI means AI algorithms are creating more data than ever. To harness the power of AI, organizations must invest in scalable infrastructure and robust data handling to manage this massive influx of data effectively.

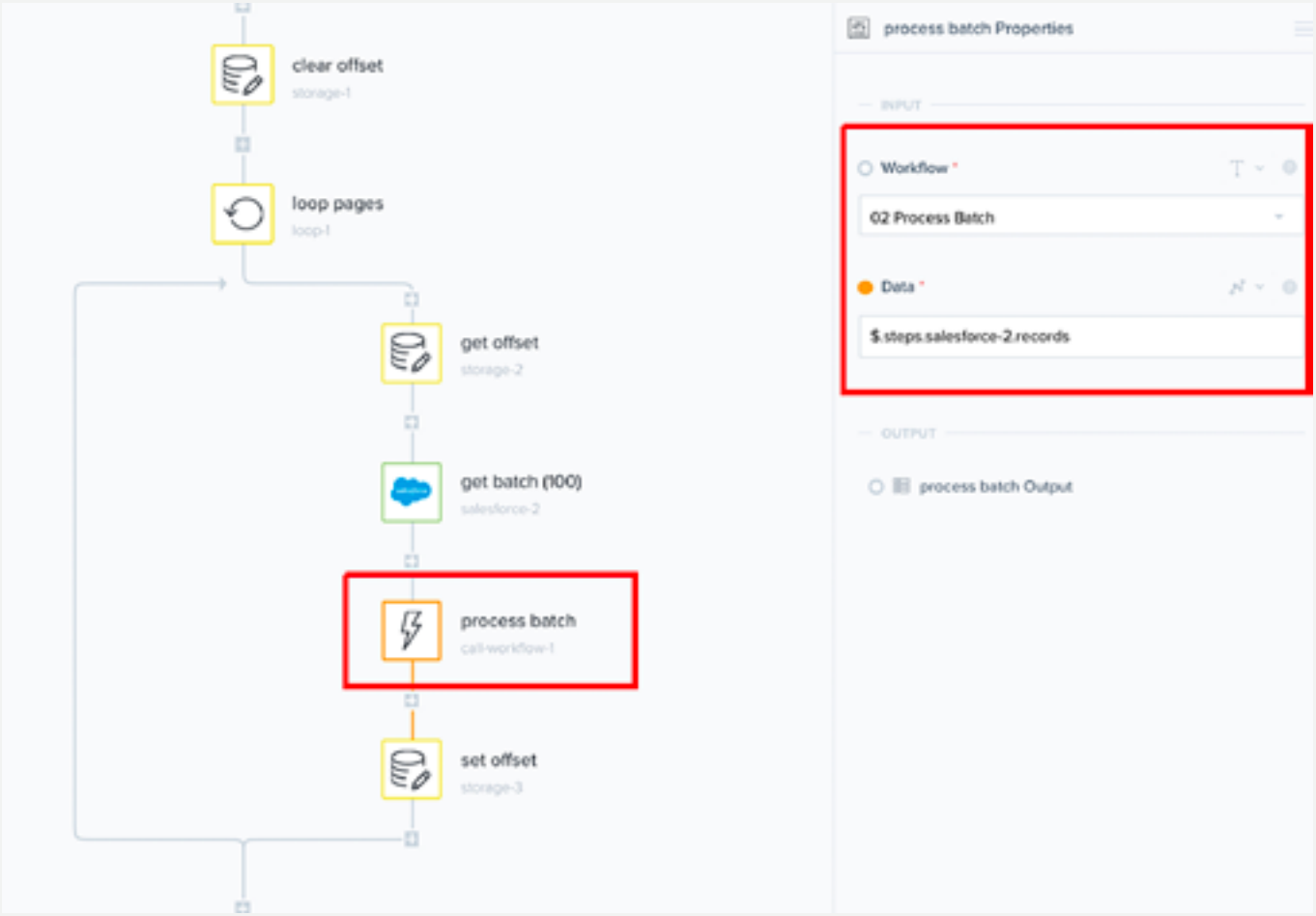
Because serverless architectures can elastically scale compute on-demand, they can efficiently massively parallelize the processes to meet growing data needs. For example, an overarching workflow can batch off chunks of data to call separate sub-workflows that operate in parallel for processing. As a result, these data processing/enrichment workflows run concurrently, processing multiple batches simultaneously, with no need to wait until the previous batch is finished.

The following diagram is a good illustration of how this works.



Serverless integration architectures enable practically unlimited concurrent bulk processing.

The contrast with less elastic architectures is clear. Typically, there are hard concurrency limits that limit parallel processing. While in a serverless integration architecture, the level of parallelism is practically unlimited, just a set of concurrently executing Lambda functions while enabling robust control over the business logic that determines the "batches."



Tray Universal Automation Cloud can execute a callable workflow to process bulk data concurrently.

Data integration case study: Enterprise software leader streamlines PostgreSQL data load

One of the world's largest technology companies sought to gain analytical visibility into their tens of millions of sales opportunity transactional detail. They needed to run reporting on those logs to provide visibility into its sales funnel for budgeting, forecasting, and reporting by loading into PostgreSQL.

The goal was to shift from a reporting window that took close to a full day to complete to load and refresh every five minutes.

By moving from hand-coded integration to Tray Universal Automation Cloud's flexible serverless architecture, they cut their processing window by 99% while also reducing the operational workload, saving 40 hours per week from a team of 8 (including six engineers) by reducing the need to monitor and optimize resources.

> "We'd previously had eight people doing 40 hours a week on this. This was a process that used to involve taking days to build things out in SQL, then spending hours in Excel. We got that down to five minutes a day, and it's fantastic."
>
> **— Business Operations Team**

SUMMARY

Today's integration and automation workloads require way more elasticity and flexible scaling without the associated increase in cost and operational overhead. Legacy integration architectures were never designed to meet this requirement.

Now, AI requires scalability like never before. The sheer volume and complexity of data, coupled with the need for real-time processing, necessitate a scalable infrastructure to handle the increasing demands. Modern integration platform scalability enables organizations to accommodate the growing data sources, handle high data velocity, and seamlessly scale their computing resources to meet the requirement of growing data flows across the organization.

It's time for enterprises to move to modern integration platforms that combine a serverless and highly parallel architecture, and flexible processing, without requiring IT/Ops sizing and provisioning to scale with today's and tomorrow's workloads.

ABOUT TRAY.AI

Tray.ai offers a composable AI integration and automation platform that enterprises use to turn AI into standout business performance. The Tray Universal Automation Cloud is a single, AI-ready platform that eliminates the need for disparate tools and technologies to integrate and automate sophisticated internal and external business processes. From prototype to production, with Tray.ai, the development of integrations, the delivery of intelligent apps and the integration of trusted data anywhere is fast, flexible and safe.

Paul Turner, Tray.ai

tray.ai