Serverless computing

Definition

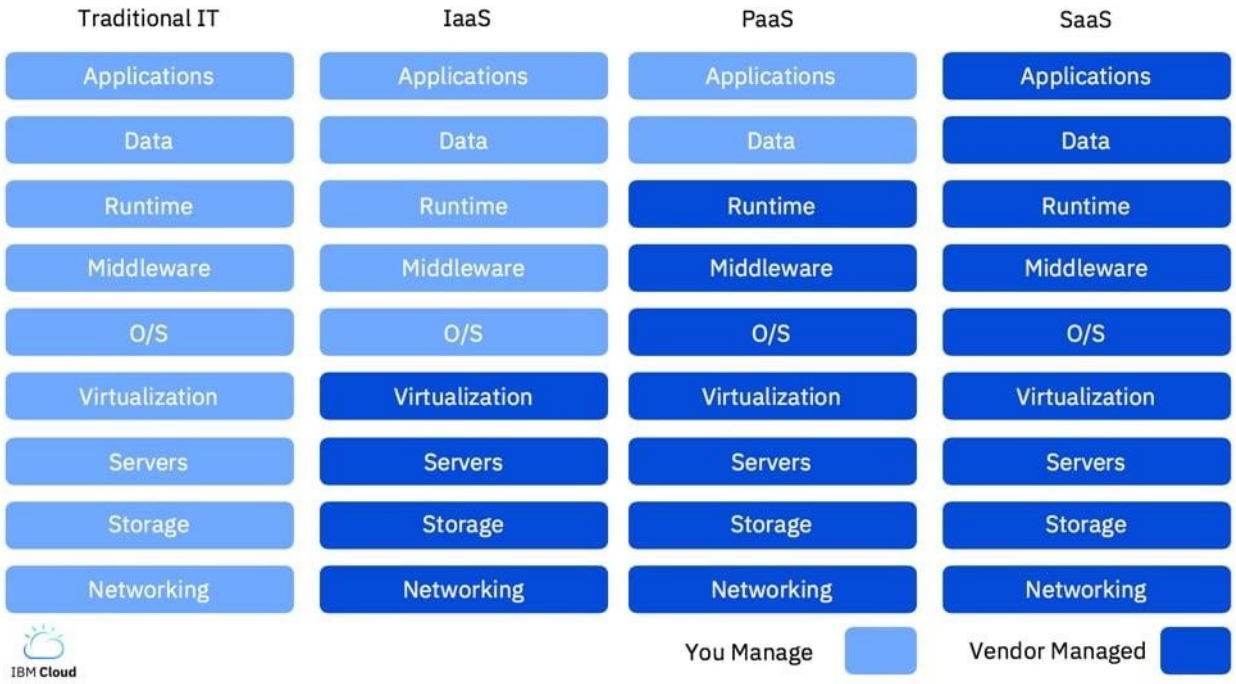
> **Serverless computing** is a cloud computing execution model in which the cloud provider runs the server, and dynamically manages the allocation of machine resources. Pricing is based on the actual amount of resources consumed by an application, rather than on pre-purchased units of capacity.

source: https://en.wikipedia.org/wiki/Serverless_computing

- Although serverless computing is arguably an oxymoron - you are still using servers to compute - the name presumably stuck because it suggests that **the cloud user simply writes the code and leaves all the server provisioning and administration tasks to the cloud provider**.

- For a service to be considered serverless, it must **scale automatically** with no need for explicit provisioning and be **billed based on usage**.

Recall: service models



FaaS + BaaS

- While cloud functions - packaged as FaaS (Function as a Service) offerings - represent the core of serverless computing, cloud platforms also provide specialized serverless frameworks that cater to specific application requirements as BaaS (Backend as a Service) offerings.
- Put simply,

serverless computing = FaaS + BaaS.

Serverless vs. Serverful Cloud

	Characteristic	*AWS Serverless Cloud*	*AWS Serverful Cloud*
PROGRAMMER	When the program is run	On event selected by Cloud user	Continuously until explicitly stopped
	Programming Language	JavaScript, Python, Java, Go, C#, etc.[4]	Any
	Program State	Kept in storage (stateless)	Anywhere (stateful or stateless)
	Maximum Memory Size	0.125 - 3 GiB (Cloud user selects)	0.5 - 1952 GiB (Cloud user selects)
	Maximum Local Storage	0.5 GiB	0 - 3600 GiB (Cloud user selects)
	Maximum Run Time	900 seconds	None
	Minimum Accounting Unit	0.1 seconds	60 seconds
	Price per Accounting Unit	$0.0000002 (assuming 0.125 GiB)	$0.0000867 - $0.4080000
	Operating System & Libraries	Cloud provider selects[5]	Cloud user selects
SYSADMIN	Server Instance	Cloud provider selects	Cloud user selects
	Scaling[6]	Cloud provider responsible	Cloud user responsible
	Deployment	Cloud provider responsible	Cloud user responsible
	Fault Tolerance	Cloud provider responsible	Cloud user responsible
	Monitoring	Cloud provider responsible	Cloud user responsible
	Logging	Cloud provider responsible	Cloud user responsible

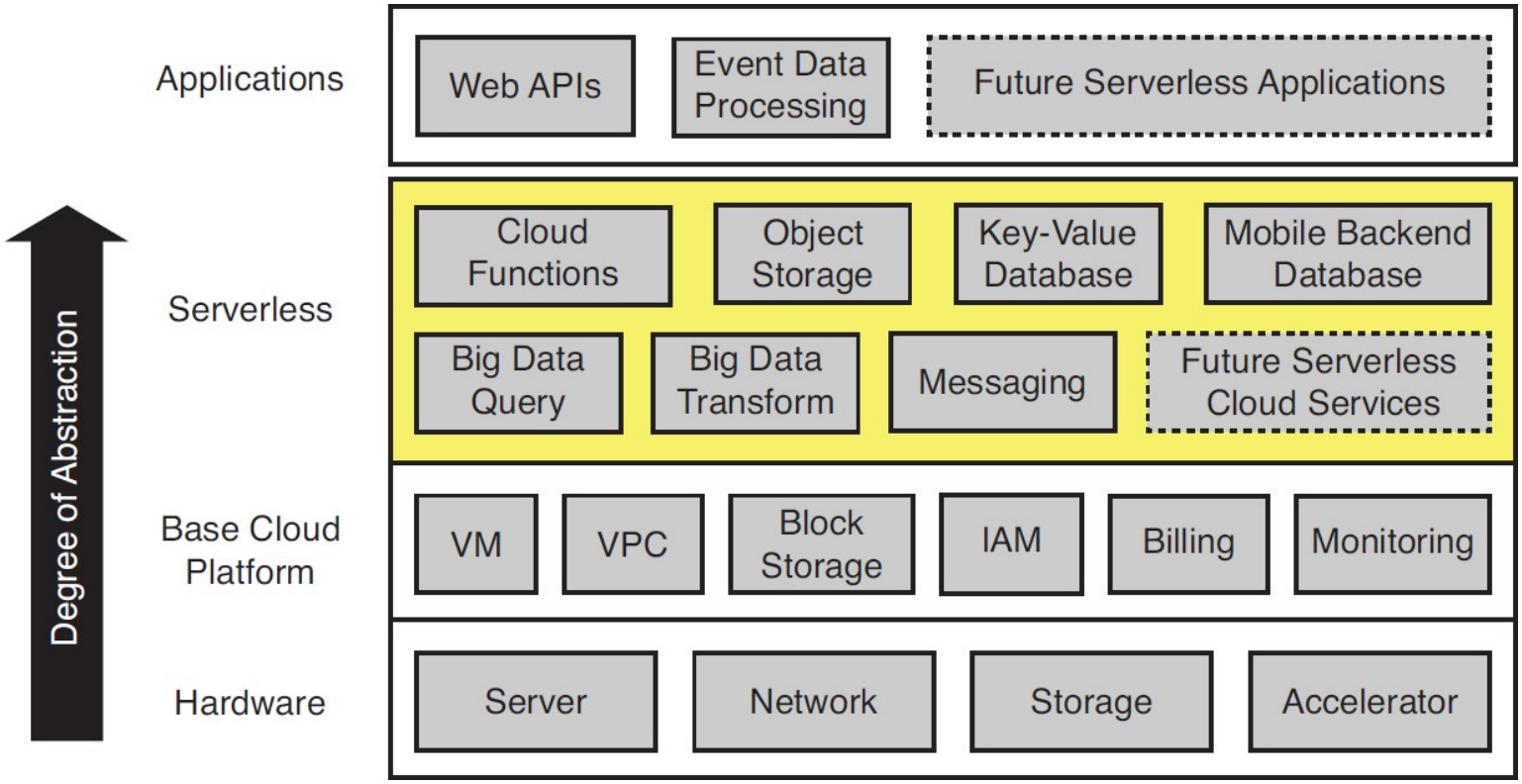
source: Cloud Programming Simplified: A Berkeley View on Serverless Computing

Serverless vs. Serverful Cloud (2)

There are three critical distinctions between serverless and serverful computing:

1. **Decoupled computation and storage**. The storage and computation scale separately and are provisioned and priced independently. In general, the storage is provided by a separate cloud service and the computation is stateless.
2. **Executing code without managing resource allocation**. Instead of requesting resources, the user provides a piece of code and the cloud automatically provisions resources to execute that code.
3. **Paying in proportion to resources used instead of for resources allocated**. Billing is by some dimension associated with the execution, such as execution time, rather than by a dimension of the base cloud platform, such as size and number of VMs allocated.

Architecture of serverless cloud



Reading list

E. Jonas et al., “Cloud Programming Simplified: A Berkeley View on Serverless Computing” (2019)

(documents available on Blackboard)

Questions, comments?