1、(a) According to definition: for a fixed workload (program size):

(The amount of execution-time on 1 server) = (The amount of execution time on n servers) * (Speedup)

$$T = S\,(\alpha T + (1-\alpha)T(n,\alpha))$$

$$\Rightarrow S = \frac{T}{\alpha T + (1-\alpha)T(n,\alpha)} \quad ①$$

(b) $① \Leftrightarrow S = \frac{1}{\alpha + (1-\alpha)\frac{T(n,\alpha)}{T}}$

As $n \to \infty$:

$$T(n,\alpha) \to 0 \Rightarrow (1-\alpha)\frac{T(n,\alpha)}{T} \to 0$$

$$\overset{①}{\Rightarrow} S \to \frac{1}{\alpha}, \quad S \leq \frac{1}{\alpha}$$

∴ The upper bound is $\frac{1}{\alpha}$,

impling that even though we have infinitely many parallel servers, the maximum speedup achieved by them is still limited by the sequential bottleneck α.

(c) For Gustafson's Law: Denote $W \to$ The amount of work that can be done by 1 server per unit time

$W' \to$ The amount of work that can be done by n servers per unit time

$\alpha \to$ sequential portion of the work (sequential bottleneck):

∴ At a unit time, # work done by n servers is calculated by: $W' = \alpha W + (1-\alpha)nW$

$$\frac{W'}{W} = \alpha + (1-\alpha)n$$

Denote speedup $S := \frac{W'}{W}$:

$$\Rightarrow S = \alpha + (1-\alpha)n$$

Zxplanation: Gustafson's law indicates that if the workload is unfixed (scaled, having many enough) then the speedup (here, the amount of work done per unit time) can be linearly and infinitely improved (ideally).

Its improvement over Amdaul's law is that it gives an appropriate and optimistic measure of the speedup of a parallel system for scaled workload, which can't be given by the fixed-workload Amdaul's law.

2. (a) Assume the RMT that can be scheduled per month to be x:

From the definition of System Availability:

$$A = \frac{MTTF}{MTTF + MTTR + 24x} = 98\%$$

$$\frac{17520}{17520 + 24 + 24x} = 98\%$$

$$\Rightarrow x \approx 13.9 \text{ hours}$$

(b) $$A = \sum_{i=k}^{3} \binom{3}{i} p^i (1-p)^{3-i} \quad ①$$

(c) $p = 0.98$: According to ①:

$k=0$: $A = 1 > 0.96$

$k=1$: $A = \sum_{i=1}^{3} \binom{3}{i} 0.98^i \, 0.02^{3-i} = 0.999992 > 0.96$

$k=2$: $A = \sum_{i=2}^{3} \binom{3}{i} 0.98^i \, 0.02^{3-i} = 0.998816 > 0.96$

$k=3$: $A = \sum_{i=3}^{3} \binom{3}{i} 0.98^i \, 0.02^{3-i} = 0.941192 < 0.96$

$\therefore$ The largest minimal number is 2

3.

Company /advances	AWS	Microsoft Azure	Google	Aliyun
Data center distributions	Currently the AWS Cloud has 81 availability zones (AZ) spanning 25 regions, with each AZ containing at least 3 data centers with 50km distance among them. AWS is also planning for 24 more AZs and 8 more regions including Australia, India, Indonesia, Israel, New Zealand, Spain, Switzerland, and United Arab Emirates. AWS also has over 218 edge locations for servers and 12 regional edge caches, ensuring low-latency.	Azure global infrastructure is made up of physical infrastructure and connective network components. The physical component contains 200+ physical datacenters, arranged into regions, and linked by one of the largest interconnected networks on the planet. Data are kept entirely within the trusted Microsoft network and IP traffic never enters the public internet.	Google Cloud's data centers provide users with 24x7 hours' services. Those data centers are spanning 28 regions, 85 zones and 146 network edge locations and are available at over 200 countries and territories. Google Cloud also has highly provisioned, low-latency network, which ensures exceptional user experience and high performance. What else, Google is trying achieving zero net carbon emissions and making energy consumptions of those data centers efficient.	Alibaba Cloud's infrastructure is built on regions (24 regions) and zones (78 availability zones). Each region has many zones and a zone consists of many scattered data centers, each of which has independent supporting facilities. Such zone has higher availability, error tolerance capabilities, and extendibility than a single data center. Aliyun also has Edge Node Service (ENS) instances, providing users with low-latency edge-computing service.
Transparency of access	AWS adds an AWS Identity and Access Management (IAM) service-linked role to the user's account for each linked service the user use. Such role gives access information using AWS CloudTrail logs, which helps monitor and audit the actions across AWS.	Azure provides Lockbox, an interface for customers to review and approve or reject customer data access requirements. Such Lockbox gives customers explicit control when certain data of them are accessed, which ensures access transparency.	Google Cloud has the Access Transparency log, which records the actions that Google personnel take when accessing the customer content. Such log ensures Access Transparency and protects the privacy of consumers	Alibaba Cloud provides Resource Access Management (RAM) and Security Token Service (STS). With them users can grant different access permissions on image resources to different users, which guarantees access transparency.
Co-location services	AWS Direct Connect enables establishing a dedicated network connection from users' premises to AWS. Also, Smartronix and AWS collaboratively created	Azure has ExpressRoute, which is a service that enables users to create private connections between Azure datacenters and infrastructure that's on	In order to provide low-latency (< 5 milliseconds) colocation facility location, Google Cloud provisions Dedicated Interconnect, a direct physical	Aliyun has Express Connect, which allows the users to establish high bandwidth, reliable, secure, and private connections between

	the Colo-to-Cloud Program, providing a solution to migrate an existing virtual machine environment for a predictable low-cost and short-timeframe migration model.	the users’ premises or in a colocation environment.	connections between users’ on-premises network and Google's network, helping connecting the users’ network to a Google Edge point of presence (PoP), enabling large-amount data transfer.	different networks based on Smart Access Gateway and SD-WAN capabilities. The service also supports peering connections between VPC networks across regions and Alibaba Cloud accounts.
Support of HPC and AI applications	AWS delivers an integrated suite of services that provides everything needed to build and manage HPC clusters in the cloud to run the most computationally intensive workloads across various industry verticals. Additionally, with access to a broad portfolio of cloud-based services like Data Analytics and Artificial Intelligence (AI), users can manipulate traditional HPC workflows to innovate faster.	Azure’s Compute (CPU/GPU based VMs), Storage, and Networking units provide the users with infrastructures to perform HPC. Azure offers management platforms like Azure Batch and Azure CycleCloud. Also in Azure, N-series VMs have NVIDIA GPUs designed for compute-intensive or graphics-intensive applications including artificial intelligence (AI) learning and visualization.	Google Cloud’s flexible and scalable offerings support scalable workloads includes the latest Intel and AMD processors, NVIDIA GPUs, and high-throughput, low-latency object and file storage. Recently Google announced the public preview of the HPC VM image, making it easy and quick to instantiate VMs that are tuned to achieve optimal CPU and network performance on Google Cloud.	Alibaba Cloud Elastic High Performance Computing (E-HPC) provides an end-to-end public cloud service. It uses Industry-leading Skylake CPU and Pascal GPU. E-HPC can eliminate job queue times and automatically scale nodes and specifications to suit business needs. Aliyun also provides HPC software to cover most user needs at multiple levels, including operating systems, task scheduling, and industry applications.
Concurrency control	AWS uses AWS Step Functions to control concurrency in the users’ distributed systems, which helps users avoid overloading limited resources in their serverless data processing pipeline or reduce availability risk by controlling velocity in their IT automation workflows.	The database engine in Azure Cosmos DB supports full ACID (Atomicity, Consistency, Isolation, Durability) compliant transactions with snapshot isolation. All database operations are transactionally and concurrently executed within the database engine that is hosted by the replica of the partition.	Google Cloud Run provides a concurrency setting that specifies the maximum number of requests that can be processed simultaneously by a given container instance. Google Cloud also provides every object with two associated positive integer fields used for concurrency control.	Alibaba Cloud provides the concurrency control (CCL) feature to ensure the stability of ApsaraDB RDS MySQL instances in case of unexpected request traffic and resource-consuming statements. Aliyun also provides DBMS_CCL package for the convenience of using CCL.

Failure management	Amazon RDS uses multi-Availability Zone, the process of providing synchronous replica of the database and switching to it at downtime, to ensure robustness. Also, AWS Cloud Volumes ONTAP allows users to easily create secondary copies of their on-premises deployments, and make sure that if one site fails, they can failover and failback between copies with no data loss. What else, AWS has Data Loss Prevention (DLP) for preventing data loss caused by viruses, malware, power failures or insider threats.	Azure provides Failure mode analysis (FMA), which is a process for building resiliency into a system by identifying possible failure points in the system. In Azure's data center architecture, Azure sets up fault domains, upgrade domains and availability set assignments for fault tolerance.	Google Cloud's HA (High Availability) provides data redundancy, which is made up of a primary instance and a synchronized standby instance. It utilizes failover (means making a standby instance the new primary instance in the event of an instance or zone failure) and failback (means fail overing to the original instance when the standby instance is down) mechanisms for failure management.	Aliyun uses SLB (Server Load Balancer, including Layer-4 SLB and Layer-7 SLB) system, which can synchronize sessions to protect the ECS instances from single points of failure (SPOFs). Also, session synchronization protects persistent connections from being affected by server failure in the cluster.
Performance enhancement	AWS has EC2 C5/C5d and M5/M5d instances, which are built on the Nitro system. This collection of AWS-built hardware and software components enables high performance, high availability, high security, and bare metal capabilities to reduce virtualization overhead.	Azure announced a set of major performance improvements for Azure SQL Managed Instances, including better transaction log write throughput for general purpose and business critical instances and superior data/log IOPS for business critical instances, etc.	NetApp and Google Cloud have partnered to offer Cloud Volumes ONTAP, a data-management layer that runs on Google Cloud infrastructure to enable enhanced control, data protection, mobility and agility for business application data.	Aliyun provides Performance-enhanced instances of ApsaraDB for Redis Enhanced Edition (Tair), which are suitable for scenarios that require high concurrency, high performance, and a large number of read and write operations on hot data.

4.

	IaaS	PaaS	SaaS
Service categories	The provider will control hardware infrastructures like servers, storage, networks, and data center fabric. The user can deploy and run VMs, guest OS and specific applications. User can also specify when to request and release the needed VMs and data.	The provider will provision a platform involving both hardware and software with specific programming tools like database, development toolkits, and some runtime supports like Web 2.0 and Java. The user can develop and deploy applications onto the platform.	The provider will provide browser-initiated application software delivered to thousands of paid cloud customers. The users needn't invest in servers or software licensing, instead they just need to pay for the software and use it.
Market share provided by AWS, Google, Azure and Aliyun	In IaaS, AWS, the leading vendor, shares nearly 50% of the market, followed by Azure having a market share about 20%, Aliyun with 10% or so, and Google Cloud with 5% or so.	In PaaS, AWS is still the leading vendor capturing about 40% market share. Then follows Azure sharing about 20%. Then comes Aliyun and Google sharing a similar amount (5% or so, with Google Cloud slightly exceeding Aliyun).	In SaaS, Microsoft Azure leads the enterprise SaaS market a worldwide market share of about 17% and such share is even increasing. Then follows Google, Aliyun and AWS sharing much less than that of Azure.
Applicability, popularity and market acceptance	Applicability: IaaS helps the users save money on hardware costs. It gives the users the flexibility to scale IT resources up and down with demand as well as keeps users from buying and managing physical servers and datacenter infrastructure. Popularity and market acceptance: IaaS is one of the fastest-growing cloud spending service with a five-year CAGR of 33.7%, exceeding that of PaaS and SaaS. IaaS is also relatively widely used in that four out of five of the companies claimed to use IaaS, exceeding that using PaaS but falling behind SaaS. IaaS, with a revenue of about $39.5 billion for 2019, relatively earns less than that of SaaS but more than that of PaaS.	Applicability: PaaS reduces time and labor to market by automating many steps in development. It simplifies users' management on multiple applications' deployment (including on load balancing, security, etc.). Popularity and market acceptance: PaaS is one of the fastest-growing cloud services, with a five-year CAGR of 29.8% by 2021, which is very closed to that of IaaS. PaaS, with two-third of the companies using, is relatively less popular than IaaS and SaaS. PaaS earns relatively less than IaaS and SaaS: It is projected $18.8 billion revenue, which ranks the least among the three services.	Applicability: SaaS helps employees and companies reduce the time and money spent on installing, managing and upgrading software. Popularity and market acceptance: Although the other delivery models are steadily gaining more ground, SaaS is still the most popular one used by 89% of the companies in the market. Also, SaaS is expected to be the top earner: The revenue from SaaS alone in 2021 is estimated to be $113.1 billion. What else, SaaS is continuing its expansion: 75% of all could workloads and compute instances will be SaaS by 2021 as predicted.

The reason why most clouds start as IaaS, then upgrade to PaaS and finally move to SaaS is that in this way most companies can first invest money on infrastructure constructions to get familiar with network, computation and storage infrastructures in advance. Then when they move to PaaS and build their platforms they can utilize the infrastructures and technologies developed by themselves, which reduces costs and makes the performance of the platforms more robust and accurate. Similarly, when they move to SaaS, they can utilize the platform that they built to develop and deploy software, which increases developing efficiency and improves the robustness of the software system. Also at this stage, the familiarity of the engineers with the platforms and infrastructures makes daily maintaining easier.

5. (a):

On resource demands: A hypervisor-generated VM may demand tens of GB for hosting the guest OS as well as the application codes, whereas a docker engine requires much less resources in that it share the machine's OS system kernel and therefore do not require an OS per application.

On creation overhead: A hypervisor-generated VM emulates a computer system based on computer architectures, including a complete guest OS, hence having a larger creation overhead. Comparatively, Docker has less overhead since it only runs its own processes, file systems and network stacks, which are virtualizations on top of a host OS.

On application isolations: A hypervisor-generated VM has full isolation from the host's OS since it creates a guest OS for each VM, which only share the physical hardware. Docker is less isolated in that it utilizes kernel-space level virtualization on a common host OS for isolation: It creates a set of namespaces for the containers, which limits the access of each container in their corresponding namespace and hence achieves isolations.

On OS flexibility: A hypervisor-generated VM has a higher OS flexibility since each VM has its own guest OS, containing a complete set of OS settings and enabling the deployment of any OS regardless of the host OS. Docker is less flexible in that the containers on a machine share a single host OS.

On host platforms: A hypervisor-generated VM is hosted by a Virtual Machine Monitor (VMM) like VMware, which is a software that manages the operations of a virtualized environment on top of a physical host machine. A Docker host, Docker Engine, is a machine which has ports exposed for querying the Engine APIs running on host OS.

(b):

CPU virtualization emphasizes performance and runs directly on the processor whenever possible. The underlying physical resources are used whenever possible and the virtualization layer runs instructions only as needed to make virtual machines operate as if they were running directly on a physical machine.

VMware supports Hardware-assisted CPU virtualization assistance (called VT-x in Intel processors or AMD-V in AMD processors), which automatically traps sensitive events and instructions, allowing trap-and-emulate style virtualization as well as providing assists to reduce the overhead involved in handling these traps. VMware ESXi also allows significant levels of CPU overcommitment (running more vCPUs on a host than the total number of physical processor cores in that host) without impacting virtual machine performance. Recently it was announced that VMware and OpenStack are supporting the virtualizations of GPUs, which makes AI and Big Data Analytics more applicable.

Memory virtualization means the process of decoupling volatile random access memory (RAM) resources from individual systems in the data center, and then aggregating those resources into a virtualized memory pool available to any computer in the cluster. Now it allows networked, and therefore distributed, servers to share a pool of memory to overcome physical memory limitations, a common bottleneck in software performance.

In the memory virtualization area, VMware developed VMkernel, a hypervisor used by ESXi. It enables the creation of a contiguous addressable memory space having the same properties as the virtual memory address space presented to applications by the guest operating system. OpenStack enables the Secure Encrypted Virtualization technology, which can be used to encrypt the memory of VMs on AMD-based machines which support the technology. Also, deploying Docker containers in one or several VMs becomes a new choice for those who needs isolated runtime environment for those light-weighted containers, hence utilizing the advantages of both virtualizations and containerizations for better and more robust service.

I/O virtualization, created by abstracting the upper layer protocols from the physical connections, virtualizes a server's multiple I/O cables with a single cable that provides a shared transport for all network and storage connections, hence making the I/O management of a server more efficient and simple.

In the I/O virtualization area, VMware and OpenStack supports Single Root I/O Virtualization (SR-IOV), which allows a single Peripheral Component Interconnect Express (PCIe) physical device under a single root port to appear as multiple separate physical devices to the hypervisor or the guest operating system, hence achieving low latency and near-line wire speed. In addition, Docker's dynamic organization of the containers' I/O architecture indicates another way of organizing I/O in addition to I/O virtualization.