ISSAI White Paper
June 2023
Cloud Computing Services versus Local Data Centers in the Age of Data

Yerbol Absalyamov (yerbol.absalyamov@nu.edu.kz), Makat Tlebaliyev (makat.tlebaliyev@nu.edu.kz), and Huseyin Atakan Varol (ahvarol@nu.edu.kz)

In today's digital world, governmental bodies, industry, academia, and nongovernmental organizations need to store, manage, and process vast amounts of data. Two popular options for this are using a *custom data center* and procuring *services from cloud providers* (e.g., Amazon Web Services (AWS), Microsoft Azure, and Google Cloud). In this white paper, we will compare the advantages and disadvantages of each option for informed decision-making.

Option 1: Computing and Data Storage in a Local Data Center

Creating a data center requires a significant investment of time, money, and resources. Here are some of the key factors to consider:

1. **Upfront Costs:** Creating a data center requires a significant upfront investment in hardware, software, and infrastructure. These costs can be substantial, and it may take several years for the initial investment to be recouped with the savings on services.

2. **Maintenance and Upkeep:** Once a data center is established, ongoing maintenance and upkeep costs must be considered. This includes costs for power, cooling, and staff to manage the infrastructure.

3. **Scalability:** Creating a data center offers greater scalability options as organizations can add capacity as needed. However, this requires additional investment and planning.

4. **Control:** Creating a data center offers complete control over the infrastructure and the data stored on it. However, this also means that stakeholders are responsible for maintaining security, technical workforce training, and ensuring compliance with regulations.

5. **Independence:** Local data center allows no external access, no need to transfer data back and forth from Kazakhstan. For instance, operations will not be negatively affected during international crises, sanctions, or accidents (e.g., damage to an undersea fiberoptic cable).

6. **National Data Lake:** Stakeholders under the leadership of the state can build a National Data Lake, where important data can be stored and used when it is needed. Governmental and semi-governmental bodies can share the data with public, private and/or research organizations based on certain access authorizations.

7. **National capacity:** If local companies and academic stakeholders can access the local data center, it will be essential to progress with research and thus will develop national

capacity. The data center can be configured to run research tasks during low usage periods (e.g., nights and weekends).

Option 2: Services from cloud providers

A range of cloud-based computing and data storage services can be procured on a pay-as-you-go basis. There are a number of big tech companies that offer these services (e.g., Amazon[1], Microsoft[2], and Google[3]). Here are some of the key factors to consider:

1. **Upfront Cost:** Procuring services from cloud service providers does not need initial capital investment since payments can be made only for services used. However, initial payment might be needed to reserve some resources for extended periods. Otherwise, quality of service issues might be faced during high demand periods.

2. **Maintenance and Upkeep**: Cloud service providers take care of maintenance and upkeep of the infrastructure, which means stakeholders can focus on their core competencies. However, this is reflected in the price of the services.

3. **Scalability:** Cloud service providers offer scalability options that allow stakeholders to add capacity as needed. This means stakeholders can easily adjust their usage based on demand. However, the price of services also elastically changes based on demand as well.

4. **Control:** Services from cloud providers means that stakeholders have less control over the infrastructure and data stored on it. Cloud service providers ensure high levels of security and regulatory compliance to protect data. On the other hand, if Kazakhstan gets sanctioned, service provider might refuse to provide services and/or lock the data. This would be catastrophic for critical operations. Another essential point is that there is no guarantee that your data will not be screened or reviewed or even worse hacked.

As listed above, both creating a data center and obtaining services from a cloud provider have advantages and disadvantages. While creating a data center offers greater control and scalability options, it also requires substantial upfront investment and ongoing maintenance costs. Cloud providers offer maintenance-free upkeep, but stakeholders have less control over the infrastructure and data. In essence, having a local data center with trained personnel would ensure some degree of independence for the Republic of Kazakhstan in the Digital World. In addition, even though the added responsibility of maintenance and operation of a data center will be cumbersome, it will also provide valuable technical know-how to the country.

At Nazarbayev University, we run NVIDIA DGX A100 servers for AI research with near full utilization. The procurement cost of each server is around 200k USD including a three-year service agreement. The data center of NU was built when the university was being built and the cost of upkeep is not very high.
In Appendix A, we provide the quotes of running our AI research on a single DGX A100 on the cloud (Amazon, Google and Microsoft). So, as of now we can say that Nazarbayev University

[1] https://aws.amazon.com/
[2] https://azure.microsoft.com/en-us
[3] https://cloud.google.com/

spent more than 2 mln USD to procure 8 NVIDIA DGX servers and some middle class servers for AI (HP, Huawei). Each year, starting from 2019, around 100 researchers of NU and several researchers from SDU and KazNU are utilizing these computational resources. What if NU or any other Kazakhstani university would rent a service from AWS, Google or Microsoft for same computational parameters, what would be the expenses in that case per year? This example shows that for Academia, where AI computational resources used almost 24/7 throughout the whole year, having its own Data center with own computational resources is more valuable, safer, efficient and cheaper. We do not transfer our data out of university, we do not really rely on circumstances out of university, we manage our resources in accordance with internal preferences and priorities, hence in house model is more efficient for us.

In Appendix B, we provide the approximate major expenses if Kazakhstan were to build a local data center locally and equip it with storage and compute resources. Taking into account the example from above, it seems that for whole Kazakhstan (Academia, State bodies, Industry) the most efficient way is to have local data center with own computational resources and storage.

Appendix A – Cost of Equivalent Compute and Storage to a Local NVIDIA DGX A100 Server

Amazon Web Services Service Prices[4]

Name	Plans	Upfront cost	Monthly/Year cost USD	Description
EC2 Dedicated instances On-Demand	On-Demand		~ 0.8 mln / ~ 9.8 mln	Advance EC2 instance (p4d.24xlarge) Enable monitoring (enabled), EBS Storage amount (16 TB), Storage amount per io2 volume (16 TB), DT Inbound: Internet (150 TB per month), DT Outbound: Internet (150 TB per month), DT Intra-Region: (150 TB per month)
EC2 Dedicated Instances Spot Instances	Spot Instances		~ 0.8 mln / ~ 9.8 mln	
EC2 Dedicated Host Compute Savings Plans	Savings Plans	~ 2.7 mln	~ 0.06 mln / ~ 3.5 mln	
EC2 Dedicated Hosts Instance Savings Plans	Savings Plans	~ 3.9 mln	~ 0.06 mln / ~ 4.7 mln	
		Total	**~1.7 mln / 27.8 mln**	

Google Cloud[5]

Name	Plans	Upfront cost	Year cost USD	Description
A2-ultragpu-8g			~ 0.3 mln	GCP instance (Nvidia Tesla A100 80GB GPU) 8 GPUs 640 GB HBM2e 96 vCPUs
n1-node-96-624			~ 0.15 mln	
Cloud Storage			~ 5.9 mln	

[4] https://calculator.aws/#/addService?nc2=pr
[5] https://cloud.google.com/products/calculator

		Total	**~ 5.9 mln**	1360 GB Bundled (3 TB) Local SSD: 24x375 GiB _____ GPU: 8 NVIDIA_TESLA_V100, vCPU 96, RAM 624GB, Total cores 56, Sustained Use Discount applied (30%) _____ **Total Amount of Storage: 102,400 GiB** **Class A operations: 1,000,000 million** **Class B operations: 1,000,000 million** **Inter-region Egress - Europe: 102,400 GiB** **Multi-region Egress - Europe: 102,400 GiB** **Intra-region Egress - Europe and Asia: 102,400 GiB** **Always Free usage included: No**

Microsoft Azure[6]

Name	**Plans**	**Upfront cost**	**Year cost USD**	**Description**
10 ND96amsr A100 v4	Pay as you go		~ 0.3 mln	10 ND96amsr A100 v4 (96 Cores, 1900 GB RAM) x 31 Days (Pay as you go), Linux, (Pay as you

[6] https://azure.microsoft.com/en-us/pricing/calculator/

10 ND96amsr A100 v4	Pay as you go 1 (year)	~ 1.8 mln	~ 3.8 mln	go); 10 managed disks – P80, LRS - 9999 GB; Inter Region transfer type, 9998 GB outbound data transfer from East US to Poland Central _____ Premium Block Blob Storage, Hierarchical Namespace, ZRS Redundancy, Hot Access Tier, 100 TB Capacity - Pay as you go, 10,000 x 10,000 Write operations, 10,000 x 10,000 Read operations, 10 x 10,000 Iterative Read operations, 10 x 100 Iterative Write operations, 1,000 GB Data Retrieval, 1,000 GB Data Write, 1,000 Index, 10,000 x 10,000 Other operations
Storage			~ 0.02 mln	
		Total	**~ 4.1 mln**	

Appendix B. Capital Expenses for Building a Local Data Center

The cost of a Tier-3 data center with 100 NVIDIA DGX A100/H100 GPUs will vary depending on the size and configuration of the data center, as well as the cost of electricity and cooling. Our analysis by collecting expert opinions in the web and reading dedicated materials shows that the initial capital expenses will be between $100 million and $300 million for a Tier-3 data center with 100 NVIDIA DGX A100/H100 GPUs (200 regular nodes and several Petabytes of storage). The following are some of the key factors that will affect the infrastructure cost of a Tier-3 data center with 100 NVIDIA DGX A100/H100 GPUs:

- **Size of the data center:** The larger the data center, the more expensive it will be to build and operate.
- **Configuration of the data center:** The type of servers, storage, networking equipment and engineering communications that you use will also affect the cost of the data center.

Operational costs will depend on the following factors:

- **Cost of electricity:** The cost of electricity will vary depending on location and the current prices.
- **Cost of cooling:** The cost of cooling will also vary depending on your location and the current climate conditions.
- **Cost of personnel:** The cost of personnel responsible for operation of Data center and its maintenance.

In addition to the infrastructure costs, you will also need to keep in mind expenses for software licenses, maintenance, operational costs and security. Plus, one should keep in mind that with each generation, the performance of the AI supercomputers increase many-fold (see Figure 1). On average, equipment becomes obsolete every three years. Therefore, upgrade and replacement should be periodical.

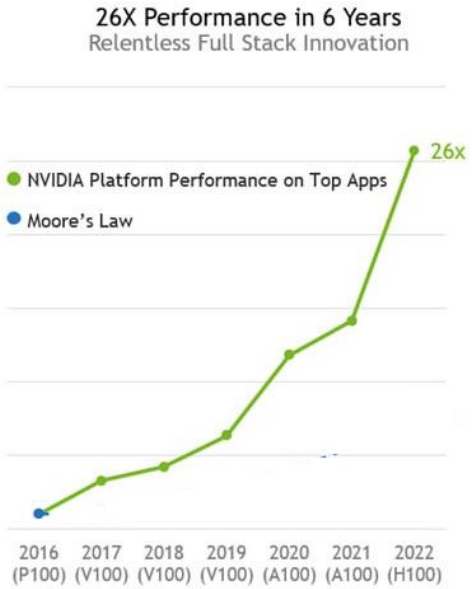

Figure 1 – The change of AI supercomputer performance in the last decade[7].

[7] https://developer.nvidia.com/blog/fueling-high-performance-computing-with-full-stack-innovation/

Here is a preliminary breakdown of the estimated infrastructure costs for a Tier-3 data center with 100 NVIDIA DGX A100/H100 GPUs:

- **Servers:** $50 - $100 mln.
- **Storage:** $25 - $50 mln.
- **Networking:** $10 - $25 mln.
- **Electrical equipment:** $25 - $50 mln.
- **Cooling equipment:** $10 - $25 mln.
- **Software licenses:** $5 - $10 mln.
- **Security:** $5 - $10 mln.

And once again we would like to mention key considerations for building a Tier-3 data center with 100 NVIDIA DGX A100/H100 GPUs:

- **Power:** Need to have a dedicated power source for your data center. This power source should be able to handle the high demand for electricity that the NVIDIA DGX A100/H100 GPUs will generate (our estimation is that at least 0.5 MW power is needed).
- **Cooling:** Need to have a robust cooling system in place to keep your data center at a safe temperature. The NVIDIA DGX A100/H100 GPUs generate a lot of heat, so you will need to have a system in place which will cool down the data center.
- **Security:** Need to have a strong security system in place to protect your data center from unauthorized access. The NVIDIA DGX A100/H100 GPUs are very powerful, so they are a target for hackers. Also, the facility needs physical safety and security.

Sources used:

https://www.nvidia.com/en-us/data-center/dgx-a100/

https://www.nvidia.com/en-us/data-center/dgx-h100/

https://www.nvidia.com/en-us/data-center/products/

https://www.nvidia.com/en-us/data-center/

https://www.datacenterfrontier.com/featured/article/11427517/nvidia-new-hardware-will-transform-data-centers-into-ai-factories

https://seekingalpha.com/article/4522089-nvidia-time-to-buy-the-king-of-data-centers

https://www.nextplatform.com/2022/05/26/datacenter-becomes-nvidias-largest-business/

https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html