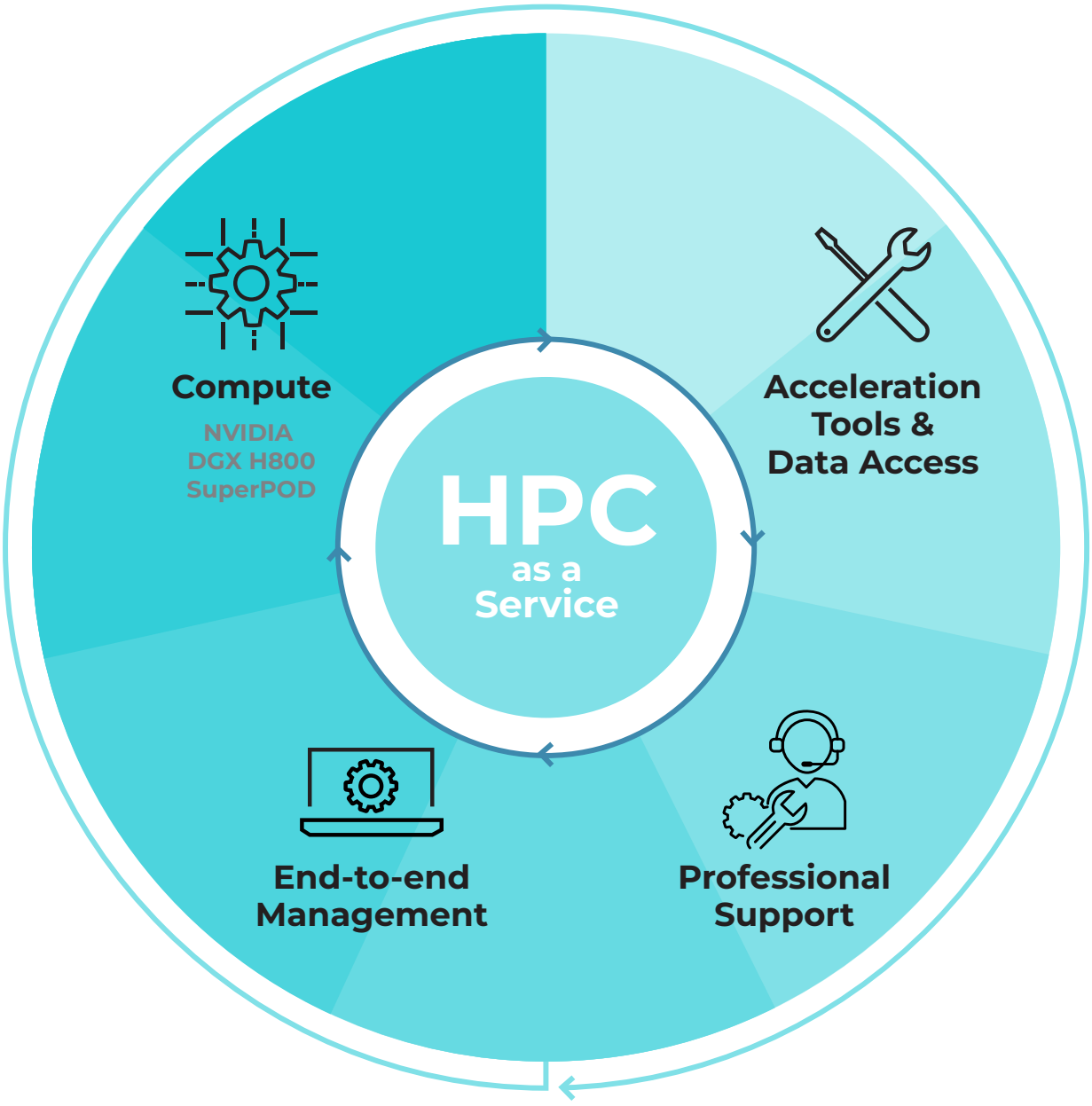
High Performance Computing

as a Service

SERVICE OVERVIEW

With HKSTP High-Performance Computing as a Service (HPC as a Service), you can innovate faster and make every dollar count.

Our comprehensive services are powered by top-notch computing, acceleration tools, and data access, providing a robust foundation to accelerate your development cycle.



SPEED

Beat Time to Market

Our services deliver top-tier computing for training, inference, and the most challenging AI workloads.

This compute foundation offers unparalleled performance and adaptability to accelerate data science pipelines.

How fast is H800? Let's compare:

Fastest Time to Train on Every Workload

A100 — 168 hours

H800 — **20 hours**

MOE with 395Bc

Enhanced Real-Time Inference Capability

A100 — 440ms

H800 — **100ms**

GPT-J 6B (Time to first token)

Source: Nvidia
The results are dependent on simulation and displayed for reference only.

VALUE

Best Value of Investment

GPU's parallel processing capability splits complex computing tasks into thousands smaller tasks that run simultaneously. This keeps your resources and team busy, not idle or wasted.

Maximise your investment value when you can expect faster results in limited time.

75%

COST SAVED

How does it make sense from a financial standpoint?

We can use MosaicML's analysis on training a 7B parameter LLM with 134B tokens.

GPU	GPU Hours to Train	Approx. Cost (HKD)	
8 x A100	11,462	1,283,744	
8 x H800	**2,000**	**320,000**	**75% cheaper**

FlashAttention2 Training on a 175B LLM

The figures for A100 references Ori, while the estimated cost for H800 is HKD20 per hour.
The results are dependent on simulation and displayed for reference only.

EXPERIENCE

End-to-End Management

Time is luxury in R&D. Your focus should be on core tasks, not cumbersome operations.

What We Offer

1 Security

2 Multitenancy Architecture

3 Acceleration Tools

4 Full-stack Environment

What You Get

- Secure and fast data transfer through a local data center managed and operated by HKSTP.
- GPU re-allocation or scale on demand at ease.
- Pre-trained AI models, software tools, and cross-industry data access.
- Workload performance intelligence, software licensing, compilers and libraries

PRICE & PLANS

We offer a straightforward bill rate.

This pricing transparency ensures you know the exact costs and adjust resources as you grow.

H800 Resource	Daily (HKD)	Monthly (HKD)	Annual (HKD)
1/4 card	-	$3,600	-
1/2 card	-	$7,200	-
1 card	$1,440	$14,400	-
2 cards	$2,880	$28,800	-
3 cards	$4,320	$43,200	-
4 cards	$5,760	$57,600	-
5 cards	$7,200	$72,000	-
6 cards	$8,640	$86,400	-
7 cards	$10,080	$100,800	-
8 cards	$11,520	$115,200	-
Dedicated DGX (8 cards)	-	-	**Contact Us**

Special Offer for Science Park companies

✓ **7 days extra** for each monthly subscription; and,

✓ **1 month Tasting Period for free***

*Note
- Each eligible company will receive 1/8 H800 card.
- Please contact account manager for application.
- HKSTP reserves the right to amend, suspend and terminate the Special Offer and its terms and conditions at any time at its sole discretion without prior notice.

FAQs

I. General

1. **What is your HPC service?**
   Our HPC service is a turnkey AI solution comprising AI Models (A), Big Data (B), and Compute (C).

   Powered by NVIDIA DGX server (H800), our cloud platform provides instant access to graphics processing units (GPUs) on demand. Users can harness powerful compute resources to accelerate streamline AI development, without the need for hardware ownership or maintenance. HKSTP manages all data centre infrastructure.

2. **What GPU options are provided?**
   HKSTP provides NVIDIA DGX server (H800) as the primary GPU option, equipped with powerful computing capabilities and interconnect solutions including InfiniBand and NVLink.

   Note: Up to 9X faster training speeds can be achieved with the NVIDIA DGX server (H800) compared to the A100 GPU, specifically demonstrated on training MOE with 395Bc.

3. **What workloads are suitable for your HPC services?**
   Our HPC services are best fit for compute-intensive tasks, such as:
   - Machine learning and deep learning model training
   - Scientific computing and simulations
   - Generative AI
   - Natural language processing
   - Deep Learning Recommendation Models

4. **Why choose our HPC service?**
   - Safe and Secure: HKSTP manages and oversees all data centre infrastructure in Hong Kong.
   - Scalable: Pay for what you use; easily adjust GPU resources to match your needs.
   - Accessible: Access powerful GPU resources from any location with an internet connection.
   - Resourceful: Utilize different pre-trained models, software tools, and diverse industry data to boost workload performance.
   - Cost-Effective: Avoid upfront expenses associated with hardware and software purchase, as well as maintenance.
   - Personalized Support: Receive professional technical assistance and consultation services from dedicated experts at our Service Centre.

5. **Who can subscribe?**
   We welcome startups, AI companies, research centres, SMEs, organisations, and enterprises.

6. **Are there any subscription capacity limits?**
   No.

7. **When can I subscribe?**
   Anytime. Subscriptions are open throughout the year.

8. **What support options are available?**
   We offer onboarding support through online documentation and professional services, along with Service Level Agreements (SLA) for technical resolution.

II. System & Security

1. **What additional tools are offered?**
   Our HPC service integrates with the NVIDIA NGC Catalog, allowing access to NVIDIA AI Enterprise, which provides easy-to-use microservices for optimised model performance.

   Additionally, users have access to a variety of tools, including:
   - Deep learning frameworks (e.g., TensorFlow, PyTorch, Keras)
   - Scientific computing libraries (e.g., NumPy, SciPy, Pandas)
   - Programming languages and IDEs (e.g., Python, R, Jupyter Notebook)

2. **Is GPU virtualization available?**
   Yes, we support NVIDIA MIG (Multi-Instance GPU), enabling the division of a single GPU into smaller isolated instances. Each instance is fully independent with dedicated high-bandwidth memory, cache, and compute cores.

   Benefits:
   - Quality of Service: Users can cater to a wide range of workloads, ensuring quality for each.
   - Cost-Efficiency: Partitioning GPUs into multiple instances allows for more cost-effective and granular GPU resource allocation, benefiting users with varied workload requirements.
   - Flexibility: Dynamic allocation of GPU instances based on demand provides scalable and flexible GPU provisioning. Users can select the MIG configuration that aligns best with their workload needs.

3. **How can I access GPU instances?**
   Access methods include:
   - Web-based console or dashboard
   - Secure shell (SSH) connection

4. **How do I transfer data to and from your HPC system?**
   You can transfer data through the staging room facility and internet connectivity.

5. **Do I need to purchase storage?**
   Users receive 500GiB NFS storage per H800 GPU card. Additional NFS storage is available for purchase at HK$100 for 1 TiB/month.

6. **How do you protect my data?**
   As your trusted partner, we take our obligation to keep you safe seriously. Our purpose-built infrastructure has put stringent security measures in place, including data encryption, access controls, and regular backups, to safeguard your data safety and privacy.

III. Subscription & Payment

1. **What are the pricing models?**
   We offer 3 pricing models:
   - Daily plan for on-demand, ad-hoc projects
   - Monthly plan for consistent GPU usage
   - Yearly plans for longer-term projects

2. **How to subscribe?**
   Sign up for an account via OnePass. Upon approval, access the Member Portal to choose GPU instances and storage, configure software environments, and manage your usage and billing.

3. **Do I need to submit any documents?**
   We need your Business Registration for opening a OnePass account.

4. **How soon can I start using the computing power?**
   The effective date can be Immediate or Scheduled upon payment completion.

5. **My computing needs are minimal. Is there any option for me?**
   Certainly. We offer flexible subscription options from ¼ to a full node of Nvidia DGX server (H800). Consider 1/4 of an H800 card with our monthly plan for minimal needs.

6. **Can I have a longer-term subscription period?**
   Please contact us by email at dsh@hkstp.org. We offer tailored consultation and personalised packages.

7. **How do I pay?**
   We accept Visa, MasterCard, AlipayHK, and Faster Payment System (FPS).

IV. Special Offer

1. **Are there any offers dedicated to Science Park companies?**
   Committed to nurturing and supporting our ecosystem, Science Park companies* can enjoy an additional 7 days for each monthly subscription. For details, contact us at dsh@hkstp.org.

   *Note: Companies participating in HKSTP's Ideation, Incubation or tenancy programmes located at Science Park, InnoCentre and DT Hub

2. **I am unsure if your HPC service is suitable for me. Can I have a trial first?**
   We hear your needs! For Science Park companies, we offer a 1-month Tasting Period*.

   Non-Park companies are welcome to visit our Service Centre. We provide on-site experience sessions, at no cost. Contact us to arrange consultations so we can cater to your needs.

   *Note: Each eligible company receives 1/8 of an H800 card. Contact your account manager for application. HKSTP reserves the right to amend the Special Offer terms at its discretion.

3. **Are there any discounts for longer-term subscriptions?**
   For information on longer-term subscriptions, please reach out to us at dsh@hkstp.org. Our team is ready to tailor a package to meet your specific needs.

HKSTP

Service Centre,
2/F, Building 19W
Hong Kong Science Park

2629 6736

DSH@hkstp.org

https://www.hkstp.org/en/programmes/digital-service-hub/high-performance-computing

BOOK CONSULTATION

2024 NOV