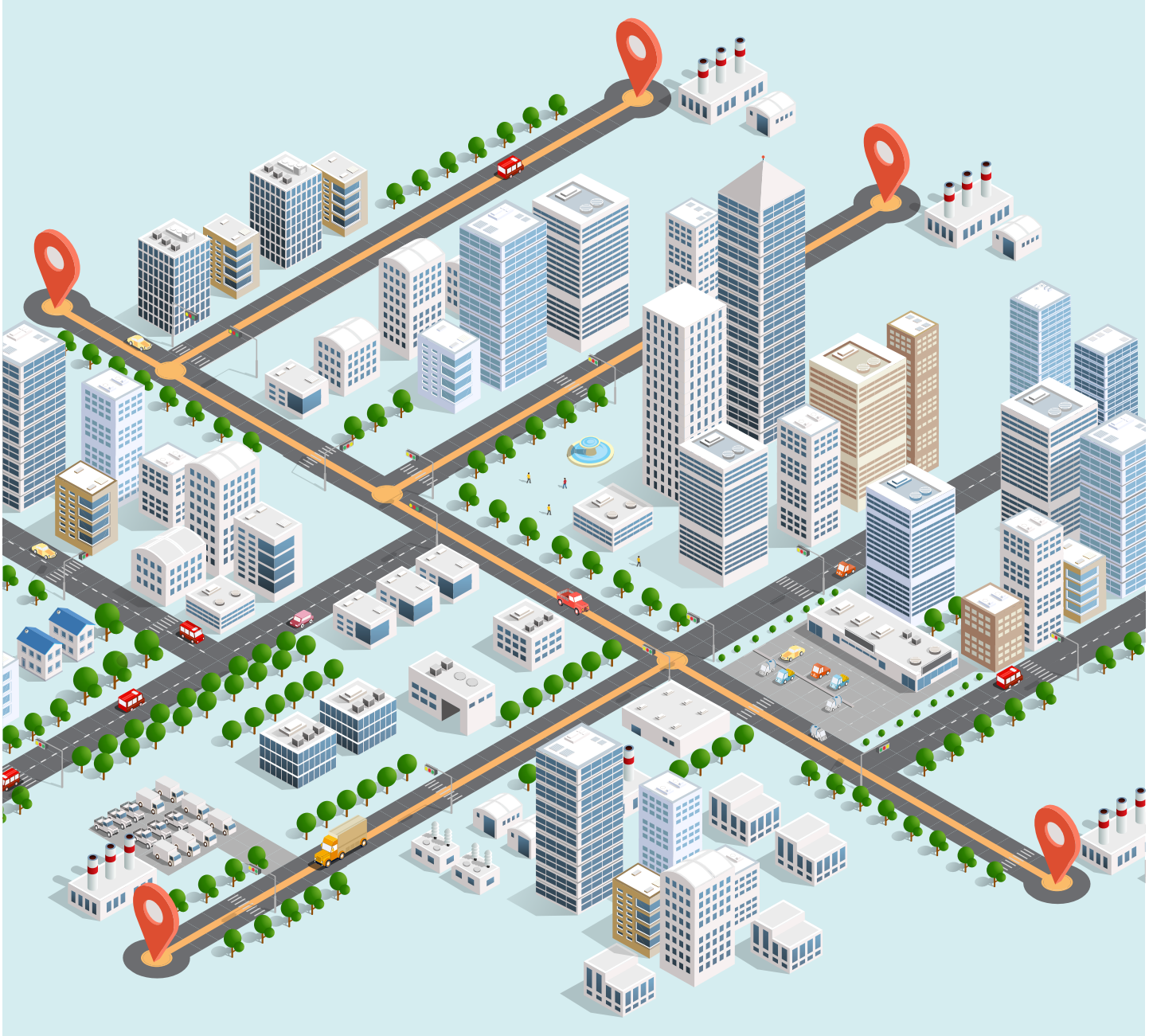
greyspark partners

OPTIMISING INFRASTRUCTURE ARCHITECTURE FOR HPC IN THE CAPITAL MARKETS

GreySpark Partners Briefing — Supported by Verne Global



Executive Summary

- Demand from financial institutions for high performance computing (HPC) capabilities grew quickly over the last decade due to a rise in the availability of technology capable of carrying out complex computations, a fall in the cost of required hardware and the greater availability of data as a consequence of the digitalisation of the capital markets. Cognizant of these trends, financial institutions began to feed structured and unstructured data into analytical systems that could optimise the performance of portfolios of securities, price exotic financial instruments, facilitate high-frequency trading and ascertain firm-wide risk. As financial institutions found ways to interrogate their data more deeply, increased demand for HPC infrastructure climbed quickly along with the associated hardware costs, despite the decreasing unit cost of the cores. Consequently, financial institutions found that the most cost-effective way to provide the required compute power was to take advantage of third-party data centre offerings and, once concerns over security were allayed, utilise, for certain non-core services, the almost limitless resource offered by hyper scale public cloud providers.

- Once financial institutions were comfortable with outsourcing their infrastructure requirements to a third-party, service providers sought to increase the stickiness of their service by offering clients added-value services that took away responsibility for additional layers of the technology stack. The proportion of the technology stack relinquished by financial institutions to third-parties varied, and a number of discrete service offerings from third-party infrastructure providers crystallised. These offerings range from:
  - rack space, power and connectivity only (colocation services); through to
  - all hardware but no software provisioned by the third-party service provider — or bare metal-as-a-service; and
  - software-as-a-service (SaaS), where a third-party provider manages on behalf of its client the entire technology stack from networking and connectivity layers all the way up to the application layer.

- Designing an ideal IT architecture for HPC is challenging in the capital markets, as the environment is highly regulated, forcing financial institutions to juggle data protection and strict levels of security with the need for scalability of compute resources. Even so, financial institutions invest in HPC because the insights it generates can deliver massive competitive advantage. Consideration of the security of the HPC environment ranks highly on the list of priorities for those deciding on their architectural approach. The decision on how high up the technology stack a financial institution will relinquish control to a third-party provider rests on not only security considerations, however, but also on the skillsets that it will need to retain in-house to achieve and maintain the cost-efficient and optimised delivery of its HPC services.

- In 2018, the maturity of the technology stack outsourcing varies from one financial institution to the next. Those feeling pressure to utilise their huge data sets better to gain competitive advantage are, typically, moving their operations into hyper scale public cloud to take advantage of the added-value services and additional infrastructure management services that come as part of the commercial package. Financial institutions, however, often need a more specialised environment for their HPC deployment than a hyperscale public cloud provider can offer. If a financial institution is using bespoke or in-house developed applications, the more tailored HPC DevOps support provided by a data centre service provider can be an advantage. Additionally, those financial institutions needing to prioritise the secrecy of their developing analytics, the option of using single-tenancy, bare metal-as-service provided by HPC specialist, industrial-scale data centres can be the most appropriate approach. For this reason, GreySpark Partners predicts that as financial institutions mature in their approach, they will increasingly begin a cloud repatriation process – moving their HPC environment from the hyper scale virtualised public clouds towards using bare metal-as-a-service.

greyspark partners

OPTIMISING INFRASTRUCTURE ARCHITECTURE FOR HPC IN THE CAPITAL MARKETS

In 2018, there is strong and growing demand from financial institutions for infrastructure that delivers the computational power needed for a range of HPC services, such as pre-trade analytics and risk management. To provision the required infrastructure is expensive, and third-party infrastructure providers — which benefit from economies of scale — are often the most economical choice. Determining which type of provider to choose depends on the answer to a number of key questions.

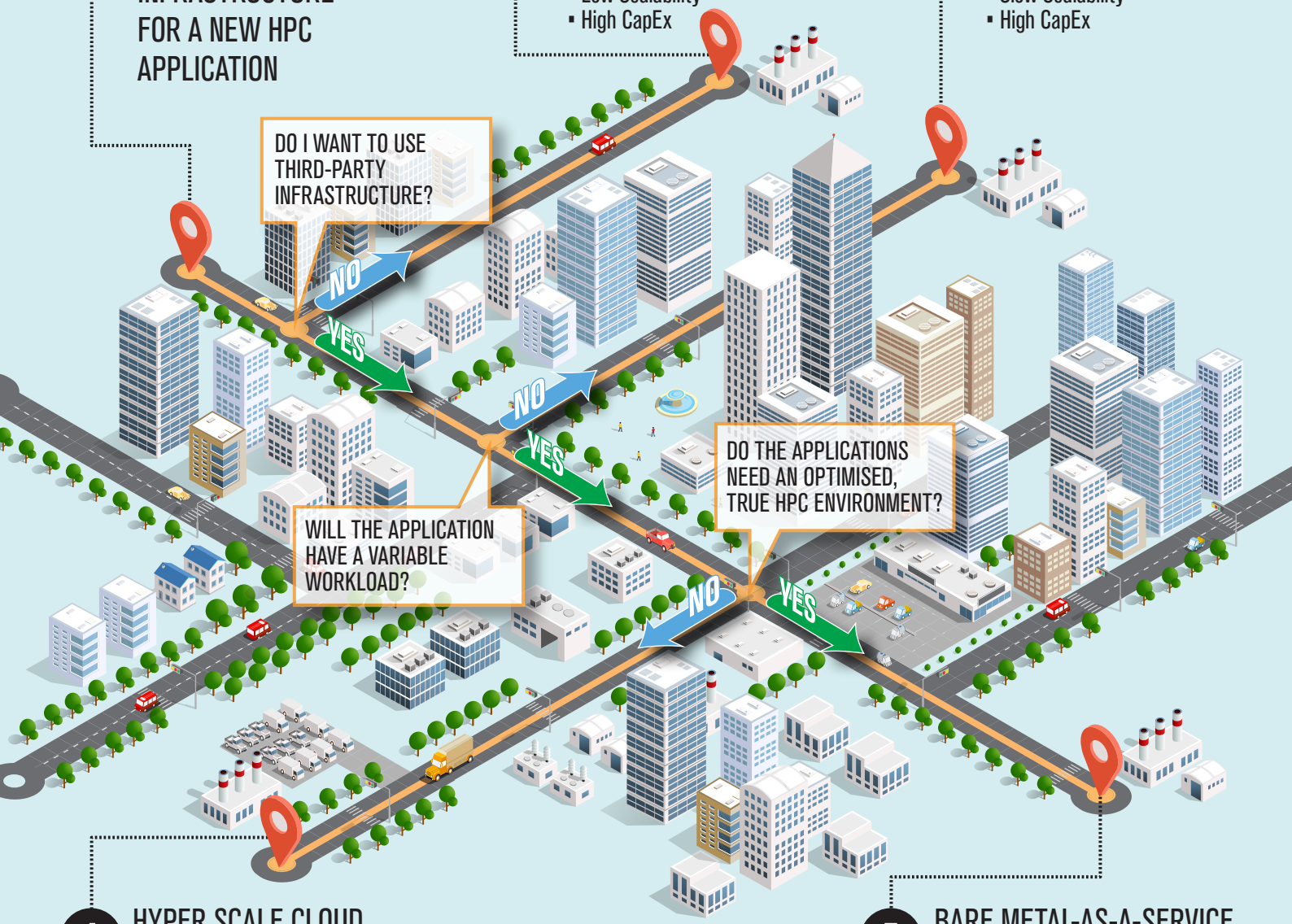
1 A FINANCIAL SERVICES FIRM NEEDS INFRASTRUCTURE FOR A NEW HPC APPLICATION

2 ON-PREMISES

- Financial Institution-controlled
- Single Tenancy
- Low Scalability
- High CapEx

3 COLOCATION SERVICES

- Hardware Flexibility
- Single Tenancy
- Slow Scalability
- High CapEx



4 HYPER SCALE CLOUD

- Simple 'Cloud-first' Option
- Multi Tenancy
- High Scalability
- Low CapEx

5 BARE METAL-AS-A-SERVICE

- Optimised HPC Environment
- Single Tenancy
- High Scalability
- Low CapEx

The maturity of the technology stack varies between financial institutions. Those firms feeling pressure to utilise better their huge data sets to gain competitive advantage are, typically, moving their operations into hyper scale cloud to take advantage of the added-value services and additional infrastructure management services that come as part of the commercial package. However, issues with multi-tenancy and the inability to stipulate the architecture in the cloud — as well as the loss of performance caused by virtualised infrastructure — means that, for HPC, financial institutions are increasingly using bare metal, single-tenancy services provided by mega-scale data centres.

For more information please visit:
greyspark.com

greyspark partners

1. HPC in Financial Institutions

As HPC capabilities are a matter of competitive advantage for financial institutions, the decision of which third-party data centre infrastructure provider to use to support applications requiring HPC is a critical, strategic decision. Financial institutions, many of which operate their own infrastructure in co-located premises or on their own premises, may find that, when it comes to HPC, retrofitting and HPC-optimising a legacy data centre is cost prohibitive.

HPC is used by financial institutions for risk modelling and quantitative research, pricing and derivatives trading, and new technologies including big data and artificial intelligence (AI) and machine learning can be used to enhance the quality of the insights generated by those processes and workflows.

Risk Modelling & Pricing – Computationally challenging quantitative models for risk management and pricing are used to estimate value-at-risk (VaR) and employ techniques such as historical simulation and Monte Carlo simulation. The latter is used to value and analyse instruments, portfolios and investments by simulating sources of uncertainty affecting their value and then determining their average value over the range of outcomes. Historical VaR is calculated by plugging in a set of historical price changes to a current portfolio to yield a distribution of changes in its value and from these the percentile – or the VaR – is determined. These two methods involve intensive computations that can be undertaken in an HPC environment.

Derivatives Trading – The increasing use of credit derivatives such as single-name credit default swap (CDS) and the creation of more complex structured credit products requires extensive use of Monte Carlo simulation methods. Pricing techniques in the credit derivatives market are often more complex than for instruments in other asset classes, and CDS markets, in particular, suffer from lower levels of liquidity, which makes appropriate pricing more challenging to achieve. HPC provides financial institutions with the scalability and the capability to rapidly develop and deliver levels of pricing for existing and new types of complex derivatives.

Big Data – The marriage of big data technology with an HPC environment can create true competitive advantage for financial institutions in the form of the insights it can generate. In capital markets, there is a developing convergence in the fundamental and predictive analysis for trading, driven by big data analysis in a HPC environment. Platforms ingest and process vast volumes of alternative data in a plethora of formats, which can then be run through complex statistical models in the HPC environment to generate trading signals that would be overlooked by a less comprehensive analysis.

AI & Machine Learning – Forward-thinking financial institutions are already well advanced in projects that utilise AI, machine learning and deep learning technologies in trading, risk, compliance and research. Machine learning models can be more efficient and even more accurate than the statistical modelling methods used historically. Machine learning researchers have demonstrated that they can reduce computing resources and energy consumption by orders of magnitude, while improving accuracy and lowering computational latencies.

With the demand for computation rising steeply across the entire enterprise, IT teams within financial institutions are challenged to provision sufficient infrastructure, ahead of surging demand. For this reason, they are increasingly turning to third-party providers, and a range of architectural options are eventuating across the industry.

2. Infrastructure Options

HPC applications hosted on massively parallel supercomputers consist of concurrent programs designed using multi-threaded, multi-process models. For HPC to most efficiently take advantage of the hardware, servers must be dedicated and closely connected to maximise the speed, power and performance available and, thus, virtualised servers are not the optimal choice for some highly intensive and industrial scale HPC workloads. Indeed, the architecture for HPC applications is complex and must be carefully crafted and optimised to achieve the best, true HPC performance.

On-premise Infrastructure – Traditionally, financial institutions were conservative about outsourcing application or infrastructure architecture, and they managed the entire technology stack. However, while the rest of the technology industry took the evolutionary step toward cloud and adopted not only infrastructure-as-a-service (IaaS) but also SaaS, conscious of regulatory constraints, financial institutions held back. As infrastructure demands grew, however, financial institutions found the on-premise approach to infrastructure provisioning uncompetitive and cumbersome, and they began to view third-party data centres favourably. For the last decade, many financial institutions have progressively moved their infrastructure off-premise in order to focus on their core strengths, leaving the management of the infrastructure and, increasingly, application management to technology specialists.

Hyper Scale Public Cloud Services – Providers of these services provision infrastructure and host operating systems, cluster managers and hypervisors, and they can spin up virtual servers on demand. Hyper scale public cloud is increasingly popular with financial institutions for non-core services as it alleviates the burden of infrastructure management for the institution, provides elastic and almost limitless compute potential alongside instant scalability and proffers a budgetary advantage in that infrastructure costs are transferred from capex to opex. Financial institutions must be aware that they may be subject to costly data transfer charges for taking data in and out of the cloud if they chose a provider that stipulates this type of contract. The multi-tenancy of servers that goes hand-in-hand with this type of service is not appropriate for financial institutions that require ultra-high privacy and robust data security, and the virtualised environment is not optimal for running certain advanced and intensive HPC applications across.

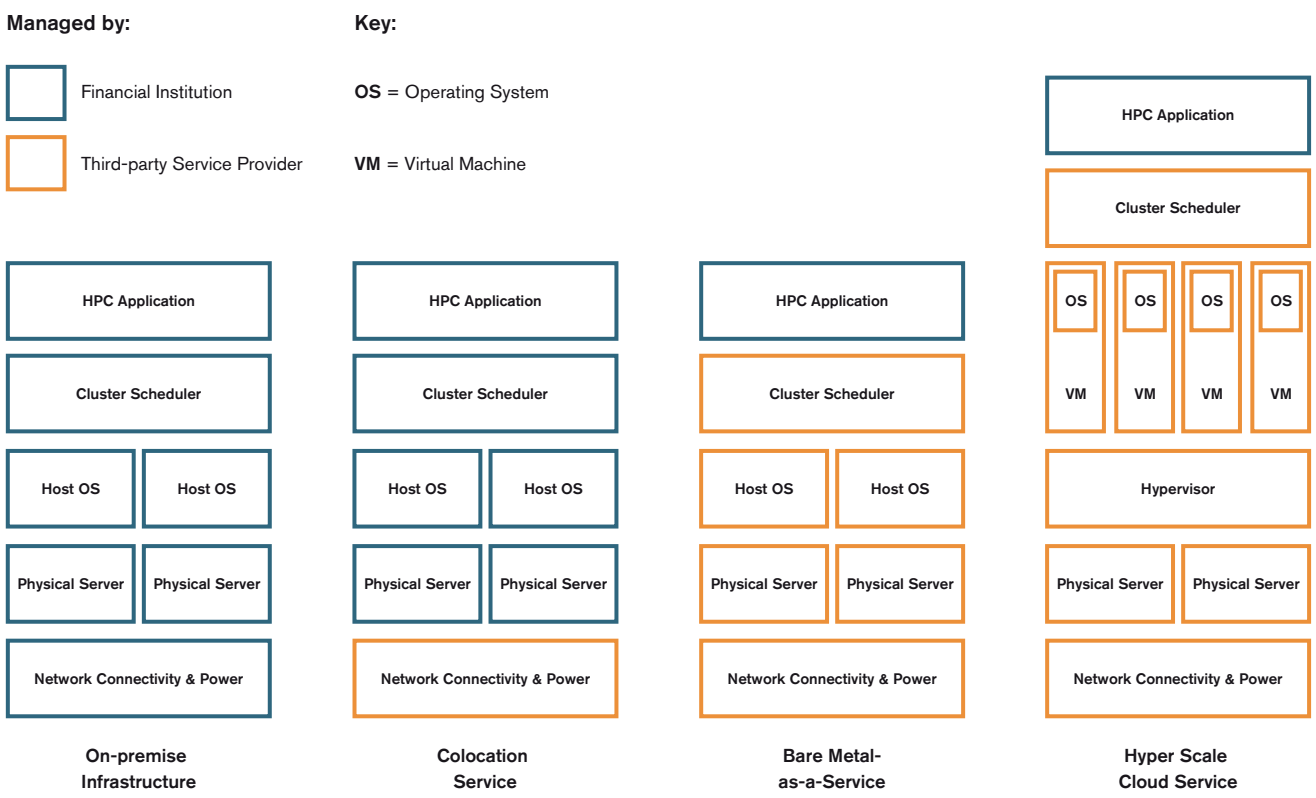
Colocation Services – Financial institutions own the devices in this operating model but rent the rack space in the data centre and pay for the power, connectivity and other services such as security on a monthly basis. If the colocation centre is located in a low-energy cost location, then this drives down the overall cost of the deployment, significantly. This option is scalable because a financial institution's compute power can grow as needed by deploying new hardware to the data centre. However, the time taken for a financial institution to provision the additional hardware may be too great for applications with a variable workload. In this case, the financial institution would bear the full cost of hardware and software required for the entire HPC service. Some specialist colocation data centres house many financial institutions and provide cross-connects where needed, as well as providing access to financial services network providers such as BTRadianz and Colt Capital Markets, for example.

Bare metal-as-a-service – Users are given access – often via a cloud portal – to the entire processing power of individual servers, as well as any storage, scheduling, networking or other services they require. Within a bare-metal infrastructure there is no multi-tenanting, and the servers provisioned are physical ones. This means that financial institutions can use their dedicated servers however they want, including running any operating system and any applications and it allows for genuinely tailored and bespoke deployments. Here, physical, dedicated servers can be provisioned to the financial institution with cloud-like ease and speed – depending on the redundancy in the data centre. Like hyper scale cloud services and colocation services, bare-metal services are typically charged for on a monthly reserved basis – such that the financial institution's infrastructure costs move from capex to opex. Service providers may offer an 'on-demand' element to enable burst capability during peak demand.

The four models are distinguished by the ownership of the hardware (see **Figure 1**).

2.0 / INFRASTRUCTURE OPTIONS

Figure 1: Four HPC Infrastructure Technology Stack Options
Source: GreySpark analysis



Many financial institutions are taking a hybrid approach to cloud, where by they utilise their own HPC infrastructure and only burst to the cloud when the workload rises above a predetermined threshold. This approach can deliver cost savings because, although the hardware and software are on-premise for the majority of the time, the additional infrastructure that the financial institution must provision to accommodate high demand periods is provisioned by the cloud providers, which, typically, charge on a pay-as-you-go monthly basis. As a consequence, part of the workload is managed in a multi-tenancy environment, which is unsuitable for many security-sensitive HPC applications and part is managed on premise which can be cumbersome and costly to maintain.

Running HPC over dedicated, un-virtualised servers gives a financial institution the full compute power of the nodes, which may reduce the cost as it lowers the number of servers needed. Thus, HPC is most efficiently achieved (lowest latency) using a colocated or bare metal-as-a-service approach.

3. Infrastructure Architecture for HPC in a Modern 'Mega' Data Centre

The decision to move infrastructure for HPC services off-premise is one that needs to be carefully considered. First, moving a service from on-premise infrastructure to a third-party's premises means that the financial institution relinquishes some aspects of the service's security to the third-party provider and this is clearly of prime concern to financial institutions. Second, financial institutions must determine the suitability and flexibility of the third-party provider's architectural design, must be considered both during the design phase and on an ongoing basis.

3.1 Security

Cyber security is of paramount importance to compute services providers in 2018. Attacks that can compromise systems are evermore innovative and, as such, InfoSec teams struggle to keep abreast of the latest patches and antivirus software to ensure their systems are protected. What is defined as the perimeter of an enterprise in the capital markets, however, is growing less obvious with the manifestation of as-a-service offerings. The division of security responsibility between the financial institution and the service provider depends on the third-party service used.

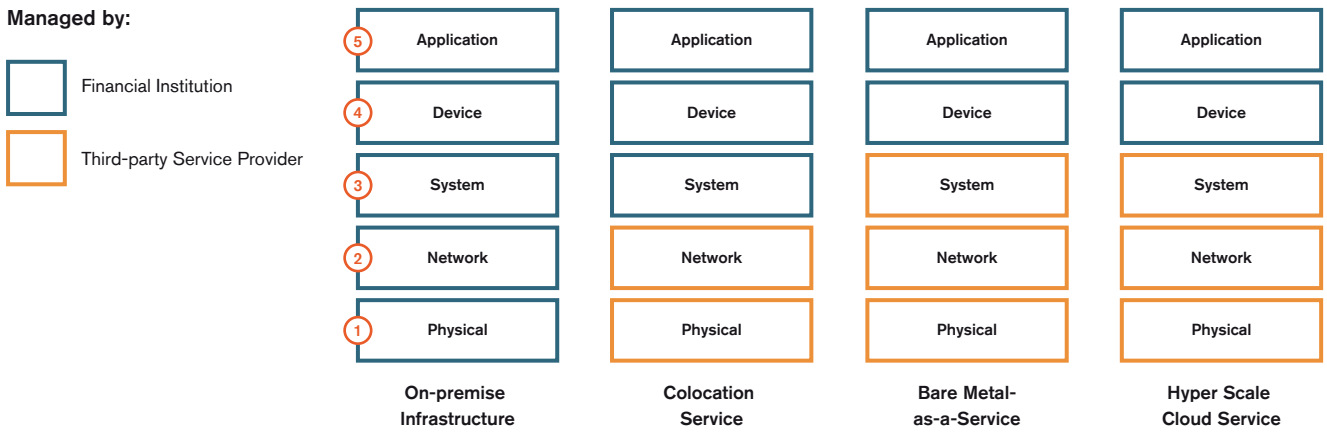
In 2013, concerns about the level of security provided by third-party infrastructure providers was a major obstacle to the uptake of their services within the capital markets. However, third-party infrastructure providers have turned the tables on this and now they are perceived as more secure than financial institutions' own infrastructure. **Figure 2** shows how the responsibilities for the security of each layer of the defence-in-depth security stack is divided up between the financial institution and the third-party service provider for each service operating model.

If the need for security is of the utmost importance, then a financial institution may shy away from the public cloud. That is not to say that its infrastructure and software up to the virtualisation layer is not secure – in 2018, public cloud is thought to be more secure than some financial institutions' IT systems. However, in GreySpark's estimation, the multi-tenancy aspect of public cloud does present a small but real threat to the security of the service.

"What is defined as the perimeter of an enterprise in the capital markets is growing less obvious with the manifestation of as-a-service offerings."

3.2 / DATA CENTRE OPTIMISATION

Figure 2: Division of Security Responsibilities for the Four Operating Models
Source: GreySpark analysis



1. Physical hardening
Guards, gates, locks, port block-outs and key cards.

2. Network hardening
Firewalls, intrusion detection and prevention systems (IDS/IPS), and general networking equipment such as switches and routers configured with their security features enabled.

3. System hardening
Reducing the surface of vulnerability, which is larger when a system performs more functions.

4. Device hardening
The default security settings of PLCs, PACs, routers, switches, firewalls and other embedded devices.

5. Application hardening
Role-based access control system, which locks down access to critical process functions, force username/password logins and combinations, etc.

3.2 Data Centre Optimisation

Optimisation of a HPC bare-metal service requires focus on two aspects: First, optimising the physical infrastructure and provision, and second, the optimisation of the hardware for application deployment.

There are several steps that should be taken to ensure that the data centre's physical hardware is optimised prior to implementation:

- **Infrastructure Technology & Operational Maturity Assessment** – A benchmark by which to compare a company's attributes to other companies within a specific industry. An example of such a model is the Control Objectives for Information and related Technology (COBIT) maturity model, an internationally recognised framework for IT governance. The specific models consist of a textual description of the target state for each level.
- **Infrastructure & Workload Capacity Assessment** – An assessment of the supportability of a financial institution's applications against the proposed schema for the third-party provider's infrastructure. In essence, it aims to deliver an assessment of the availability and capacity trends that can identify pain points.
- **Infrastructure Operations Assessment** – An 'as is' plan of the infrastructure in use currently, providing the HPC service is already in existence. The third-party infrastructure provider should review the current situation and ascertain the problem points in the network and infrastructure.

3.2 / DATA CENTRE OPTIMISATION

- **Infrastructure Environmental Impact Assessment** – The heat generated by the servers in an HPC environment can be phenomenal. Even using solid state drives (SSD) the server heat output is significant. In 2008, the EU launched a Best Practice Code of Conduct for data centres, which aims to encourage data centre operators to voluntarily measure the energy efficiency of their operations, put in place action plans to increase those efficiency gains over time and to submit annual progress reports to the EU's Joint Research Centre (JRC). The financial institution's IT teams must assess the impact that proposed service would have on the environment in which the service operates.

- **Additional Total Cost of Ownership & Return on Investment Assessment** – As a move to the cloud is predicated on reducing infrastructure costs, it is important that the data centre is able to produce meaningful TCO reductions and ROI gains for a financial institution.

- **Infrastructure Target Architecture Assessment** – There may be several ways to optimise the architecture for HPC, but the transition to the new service can be simplified if the chosen approach aligns with the existing infrastructure of a financial institution.

- **Infrastructure Appliance Capacity Assessment** – Only applicable if a financial institution were to choose the colocation or bare metal-as-a-service offering to host its HPC. The assessment should include compute architecture (CPU cores, memory, clustering details), storage capacity (flash and disk drives), product features (replication, data reduction, caching, backup and disaster recovery), as well as hypervisor support – if relevant – as well as management capabilities and networking.

- **Infrastructure Designer Service** – For the best optimisation, this should be completed initially in the design phase of the implementation, but then also reviewed on a continual basis by the data centre to ensure that the service is optimised on an ongoing basis.

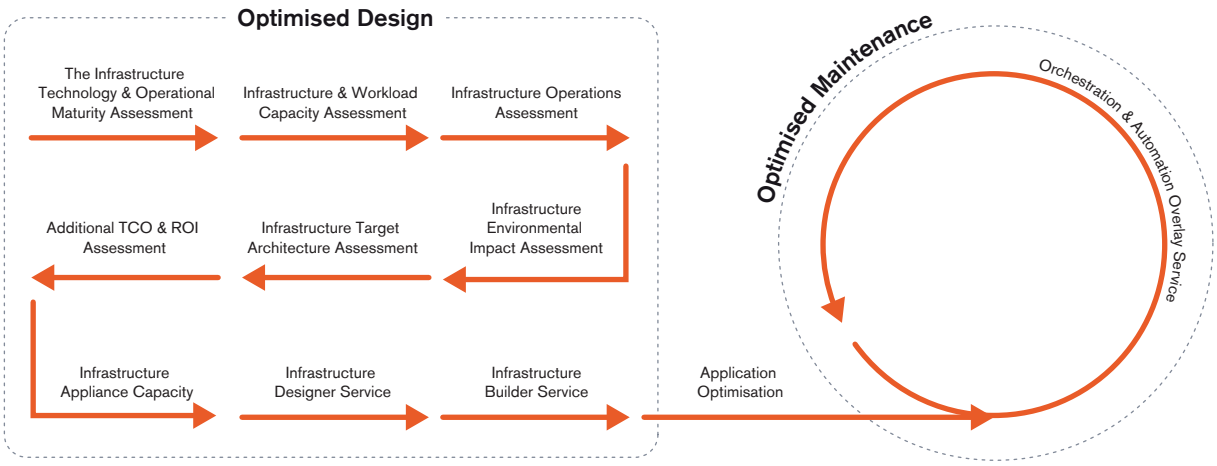
- **Infrastructure Builder Service** – Specifically, in a bare metal-as-a-service and colocation set-up, the data centre should be able to offer to build the infrastructure in situ, such that the work is done by staff within the data centre with only minimal input from the financial institution.

The optimisation of the hardware for application deployment requires the data centre to provide HPC DevOps support in order to tailor hardware to the application's requirements. This service is not included in all third-party service providers' offerings in this space and is typically absent in the more 'vanilla' generic compute hyper scale public cloud offerings.

Once the service is running, data centres should continue to work to optimise their architecture for their clients. To achieve this, the data centre and financial institution IT teams must coordinate on the following activities. An orchestration and automation overlay service is an ongoing activity that should be provided by the data centre to ensure that any variable demand in the HPC environment can be managed. The entire optimisation process in overview is pictorialised in **Figure 3**.

Bare-metal service providers have a level of control over which of the servers should be brought online as a financial institution's workload peaks. In other words, even with a bare-metal service there is some dependency on the provider to provision additional architecture. Thus, the degree to which the provider is able to optimise the network dynamically must be considered and factored into an SLA.

Figure 3: Optimising Design and Ongoing Maintenance of the Data Centre HPC Architecture
Source: GreySpark analysis



4. Making the HPC Infrastructure Decision

Although financial institutions may be tempted to engage with hyper scale, virtualised public clouds when it comes to rolling out a 'cloud first' policy, these providers, as previously discussed, may not provide an optimal HPC experience from an application perspective. Virtualised public clouds are suitable for many forms of scalable compute, but financial institutions need to remember that one size does not fit all, especially when it comes to HPC.

For small-and-mid-sized businesses that want to benefit from the features of larger data centres without the high cost associated with so doing, colocation services may be most suitable. However, for financial institutions that desire additional flexibility, tailoring and security around their HPC service, bare metal-as-a-service is the optimal solution.

The design of the system and upgrades to the hardware should be managed by the data centre service provider, which can then provide an optimised and scalable, single-tenancy service. In 2017, Forbes noted that a recent study found that 20% of cloud users had already moved one or more workloads from a public cloud environment to a private cloud. Another 10% planned to move some of their workloads from public to private cloud environments during the coming year.[1] Although these statistics are not capital markets-specific, GreySpark Partners predicts that as financial institutions mature in their approach, they will also increasingly begin a cloud repatriation process – moving their HPC environment from the hyper scale virtualised public clouds to private bare metal-as-a-service offerings.

[1] Thome, G., 2017, Elvis has Left the Building: Why Some are Exiting the Public Cloud, *Forbes* [online] Available at: <https://www.infoworld.com/article/3201104/data-center/elvis-has-left-the-building-why-some-are-exiting-the-public-cloud.html>.

GREYSPARK'S CAPITAL MARKETS INTELLIGENCE PRACTICE

TURNING EXPERIENCE INTO KNOWLEDGE.

GreySpark's dedicated analysis and research team specialises in producing high-quality industry reports, and it has extensive experience assisting industry leaders in understanding and then developing successful strategies.

GreySpark consultants have hands-on experience in bringing improvements to capital markets and have close relationships with capital markets decision-makers and industry subject matter experts. This experience allows GreySpark to maintain an overarching perspective of important changes influencing the industry enriched by a strong track record of successful consulting work that allows early observation of those changes before they become a reality.

Those observations are supported with quantitative and qualitative research, interviews and surveys. GreySpark shares those observations in industry publications, where subject matter experts comment on the impact of regulations, technology trends and shifting e-commerce business models across the industry. Those observations are also used in a client-specific context as a part of the firm's overall strategy advisory services.

For more information, please visit: **greyspark.com/research**

greyspark partners

RICH, GLOBAL IN-HOUSE LIBRARY OF **BENCHMARKS** ON RISK AND DATA MANAGEMENT, ELECTRONIC TRADING AND E-COMMERCE

ABOUT VERNE GLOBAL

Verne Global delivers advanced data centre solutions at industrial scale, allowing high performance and intensive machine learning applications to operate in an optimised environment. Verne Global's hpcDIRECT platform provides customisable bare-metal HPC, and an attuned, *True*HPC experience which is fully supported by HPC specialists.

VERNE GLOBAL

LEADERS IN HIGH PERFORMANCE COMPUTING