This research investigates “Matrix Workloads,” a workload type characterized by being able to process very large amounts of data in real-time. Examples of Matrix workloads include real-time systems of intelligence, real-time analytics, AI inferencing, robotics, autonomous vehicles, and other data-driven real-time or near real-time use cases. In this research we put forth and attempt to defend three key themes:

1/ Matrix workloads can help deploy more aggressive end-to-end data strategies, which together can greatly simplify and automate operational processes.

2/ The value of well-designed Matrix workloads together with an end-to-end data strategy can deliver an order of magnitude greater value than the traditional applications they enhance.

3/ New hardware and software architectures, techniques and tools are required to develop and deploy matrix-based workloads in the enterprise. These technologies are likely to derive from consumer technologies.

Wikibon projects that Enterprise Matrix Workloads will be 42% of Compute revenue by the end of this decade.
Premise

This research investigates “Matrix Workloads,” a workload type characterized by being able to process enormous amounts of data in real-time. Examples of Matrix workloads include real-time systems of intelligence, real-time analytics, AI inferencing, robotics, autonomous vehicles, and other data-driven real-time or near real-time use cases. In this research, we put forth and defend three key themes:

- Matrix workloads can help deploy more aggressive end-to-end data strategies, which together can significantly simplify and automate operational processes.
- The value of well-designed Matrix workloads, together with an end-to-end data strategy, can deliver an order of magnitude higher value than the traditional applications they enhance.
- New hardware and software architectures, techniques, and tools are required to develop and deploy matrix-based workloads in the enterprise. These technologies are likely to derive from consumer technologies.

Wikibon projects that Enterprise Matrix Workloads will be 42% of Compute revenue by the end of this decade.

Executive Summary

Matrix Workload Technology & Capability

We believe real-time Matrix workloads will be a significant contributor to the implementation of digital innovation and differentiation over the next decade. Also, the following four points are noteworthy in terms of understanding this emerging type of workload:

1. Real-time inference Matrix workloads are already processing in large numbers in the consumer-volume space. Examples include facial recognition, digital photography, voice recognition and enhancement, and health monitoring.
2. These technologies can enable breakthrough performance improvements relative to traditional enterprise workloads by factors of 10 to 1,000.
3. These workloads are data-intensive and will require enterprises to establish an end-to-end data strategy.
4. Disruptors can use expertise in Matrix workloads together with establishing an end-to-end data strategy to simplify operational processing and improve speed to deploy innovation radically.
5. Wikibon projects that the growth of Enterprise Matrix workloads will be robust over the second part of the decade and will represent about 42% of enterprise compute spend by the end of the decade.

We believe artificial intelligence (AI) is a relevant framework to develop real-time inferencing, which is a Matrix workload. However, other technologies such as advanced analytics, Bayesian networks, Causal inferencing, and others will also be used either together with or separately to develop real-time inference code and other Matrix workloads.

Continuing Wikibon Research into Matrix Workloads

This research is the first in a series that will define Matrix workloads and the hardware and software requirements to make these workloads possible to execute in real-time. Also, we investigate the compute types that will be necessary to deliver Matrix workloads, and the platforms, architectures, and vendors likely to drive enterprise Matrix workloads.
* I chose the term “Matrix workloads” in memory of moving from scalar to matrix operations in statistics before the electronic calculator, and in the fear and joy of tensor ranks. And of course the iconic Matrix movies, with the latest *The Matrix 4* now being filmed in San Francisco. I recently learned that David Moschella also uses the term in his book “Seeing Digital.” In it, he describes moving from a cloud of services somewhere out there to a ubiquitous matrix of intelligent capabilities, a compelling viewpoint.

## Matrix Workloads 101

### Matrix Workloads and Artificial Intelligence (AI)

“Matrix” workloads are primarily data-driven. The volume of data is very high and is often parallel by nature, i.e., sound, images, video, radar, ultrasound, and any number of IoT devices. AI is an important technology that can assist in developing Matrix workloads.

Figure 1 below shows AI development in blue on the left-hand side and AI execution in green on the right-hand side. Usually, AI development splits into data engineering, and statistical modeling and training, as shown in Figure 1. Both are storage bound batch workloads where I/O usually is the bottleneck. NVMe over Fabric NAND flash storage can offload IO overheads and allows efficient, direct connection to multiple significant data sources at a compute location.

Companies such as Pure, in particular, have developed innovative private cloud hardware platforms optimized for preparation and training, and Dell, HPE & IBM also have specialized integrated systems. As a development workload, demand for infrastructure resources is very intermittent and well suited to as-service cloud offerings.

Wikibon also points out that machine learning needs relevant quality data to be valid. Quantity without quality or relevance results in “garbage in, garbage out.”

**Bottom Line:** AI development workloads require access to quality and relevant data, and the end-to-end data architecture must ensure this is the case. However, AI development workloads are not real-time Matrix workloads.

---

### Figure 1: Relationship between AI and Matrix Workloads

*Source: © Wikibon 2020*
Real-Time AI Inferencing is a Matrix Workloads

Inference workloads are execution workloads, and the characteristics are in green on the right-hand side of Figure 1. Inferencing is usually compute-bound and needs to be close to the data sources to improve execution times. The inferencing bottleneck is usually memory bandwidth. The memory is often SRAM heavy to improve in-memory compute workflows. The compute hardware is heterogeneous (see “Heterogeneous Compute Architecture (HCA)” section below). HCAs allows a much higher degree of parallelism and speed to completion. These systems increasingly deploy neural networks and are optimized for inferencing.

The implications are far-reaching. Real-time inferencing is a matrix workload and is by far the most compute-intensive component of AI. AWS points out that over 95% of compute resources for Alexa are used for inference compute. Wikibon projects that percentage will rise to 99% by the end of the decade. AI and ML are an essential source of inferencing, but not the only source (see in the Premise section above for additional inference sources). One of the requirements of an end-to-end data strategy is to ensure the availability of good quality and relevant data for developing and improving inferencing.

Matrix Workloads Moves to the Edge

Real-time also means that compute is adjacent to equipment creating the data. There is no time to transfer data even a few miles. The cost of moving data and the reduction of data quality from loss of context are both factors that will result in a significant portion of compute moving to the Edge over the next decade. Earlier research from Wikibon shows the cost case for moving compute to the Edge is overwhelming.

Heterogeneous Compute Architecture for Matrix Workloads

Matrix workloads have large amounts of matrix data to compute. Traditional processors cannot calculate this in real-time. However, deploying many parallel accelerators with different limited instruction sets can enable completion of the workload in real-time.

Real-time Matrix workloads require specialized architectures. These are heterogeneous compute architectures.

Heterogeneous compute architectures are composed of many different compute elements (processors or cores) with varying instruction sets. These elements can include general-purpose processors, GPUs, NPUs, ASICs, FPGAs, and others. An SoC (System on a Chip) provides the interconnections to integrate everything. An outstanding example is the consumer Arm-based Apple iPhone 11 SoCs.

The significant advantages are radically improved performance and lower electrical power requirements. The primary design challenges for heterogeneous systems include managing access to memory and data and programming complexity. The CXL industry standard is emerging as a possible standard high-speed CPU interconnect that will accelerate solutions.

In conclusion, heterogeneous compute allows a much higher degree of parallelism and speed to completion and is very well suited for Matrix workloads. Wikibon expects to see a rapid increase in enterprise HCA over the next decade. The section titled “Upcoming Matrix Workload Research” below discusses Wikibon’s planned research into Heterogeneous compute.

Innovation Happens First in Consumer-volume Technologies

New technologies are usually first adopted in the consumer-volume space and are then selected by enterprises a few years later. A good example is Intel x86 consumer PC technology, which with Microsoft Windows software, became a dominant duopoly. The x86 volumes, lower costs, and advanced technology
allowed Intel to expand into enterprise servers, and replace the majority of RISC servers over a decade. x86 now dominates the compute market and a now shrinking PC market.

The introduction of the Apple smartphone in 2007 started to move expenditure to mobile devices. The sheer volume of and investment in consumer devices now drives the majority of hardware and software innovation. Arm-based systems power almost all mobile technology, and the number of Arm-based wafers created is now ten times greater than any other platform. This innovation is coming from companies such as Apple, Arm, Google, Nvidia, Qualcomm, Samsung, and many other consumer hardware and software vendors.

Consumer Matrix workloads are in full swing on consumer-mobile platforms. Apple uses facial recognition and neural cores for securing financial transactions. Google deploys a neural chip to improve image and voice processing in the Pixel 4 smartphone. Both are using neural networks to enable real-time photography and video enhancement, enhance voice recognition, health monitoring, and many other areas. Both emphasize the importance of neural networks and place much less emphasis on GPUs.

The potential of these advances has rapidly spread to mobile app developers, especially gaming apps. Large numbers of consumer developers are training themselves on the techniques and technologies required. This source of developers will be an essential source of talent as enterprises start to adopt real-time Matrix workloads.

**Bottom line:** the technologies and software used for the large and rapidly expanding Matrix workloads in the consumer space are very likely to be deployed in the enterprise space, as well as the experts.

**Case Study I - Tesla Autonomous Edge Real-time Matrix Workloads**

Wikibon believes that real-time Matrix workloads will be the foundation for the majority of high-value and effective enterprise digital initiatives. The business Edge will be a vital early implementation point for IoT initiatives. Because of MEMS**, sensors have plummeted in price, and a massive amount of IoT data is available at the Edge. Real-time Matrix workloads will enable significant simplification and automation of workflows at the Edge.

An excellent example of the successful deployment of an enterprise Matrix workload is Tesla, which is developing an Automated Driving System (ADS). The business objective of ASD is to improve the function and margins of their electric vehicles and to open up new business opportunities. This example illustrates the software, hardware, and end-to-end data implications for a real enterprise running real-time Matrix workloads at the Edge.

**MEMS (Micro-Electro-Mechanical Systems) are miniature mechanical and electro-mechanical devices made in silicon that use many of the silicon chip technologies to reduce cost dramatically over time. These microsensors can convert the analog into digital signals and are now usually more accurate than the original macro-sensors they replace. A mobile phone includes an ever-increasing number of sensors measuring acceleration, compass bearing, barometric pressure, color, GPS position, proximity, and others. The BOM price for all these sensors is about $1.50.**

**Tesla - The Dream**

The reaction time, reliability, and accuracy of ASD are potentially far better than any human driver. Matrix technologies are responsible for primary life-death decisions. In the long-run, ASD will radically reduce the 35,000 people killed in vehicle accidents every year in the US, as well as the many more accident-related severe injuries. The NHTSA study assesses the cost of road accidents is about $600 billion every year. They state, “94% of serious crashes are due to human error. Automated vehicles have the potential to remove human error from the crash equation, which will help protect drivers and passengers, as well as bicyclists
and pedestrians.” The NHTSA also states that the age of automated driving will start in 2025.

Tesla is on a journey to ASD and hopes to be first across the line. Tesla electric vehicles have independent electric engines. The brakes are separate and regenerative. There is feedback from all the wheels as a result of actions from the electric motors and brakes. The low-latency electronic components have precise knowledge of the position of the vehicle, together with knowledge of the vehicle's internal data and capabilities. This strategy creates an intra-vehicle end-to-end data architecture and creates the potential for a more responsive and much safer ride.

ASD also needs a way to integrate all the internal sensor information with external sensor data from the vehicle and create action plans. The final challenge is how to assess the quality of these plans and improve them over time. The strategic challenges are to understand the data requirements and set up an end-to-end data strategy that will create an improvement feedback loop at an affordable cost.

**Tesla Real-time Matrix Workload Problem**

The first problem to solve for the Tesla Matrix workload is how to process 1 billion pixels/second coming from the eight cameras running at 60 full-frames/second in real-time. Also, there is input from radar, 12 ultrasonic sensors, high-precision GPS, maps, and all the internal vehicle sensors such as wheel ticks, steering angle, acceleration, etc. A system needs to process these inputs in real-time, and produce a plan of action. Together, all this data drives the autopilot driver-assist and all updates to autopilot and other functions.

Tesla refers to this system as the Tesla HW3 (Hardware-level 3). Some of the constraints (such as electrical power) will not be important for many enterprises. However, we believe the end-to-end architectures and data management decisions made by Tesla will reflect the type of design that will be important to most enterprises.

**The journey to Tesla HW3**

The Tesla software philosophy is that all the data should be processed by a single system, which theoretically provides the best quality of outcome and fastest response. The original Tesla compute hardware uses MobileEye EyeQ technology, but this is too slow. The Tesla HW2 (Hardware-level 2) uses Nvidia Drive hardware and uses GPUs to process the matrix data. Tesla engineers, like the Apple & Google engineers mentioned above, understand that neural networks are much faster than GPUs to handle higher levels of data abstraction.

At the time (2016), there were no other suitable off-the-shelf hardware or software solutions available. As a result, Tesla invested in developing a hardware solution, the HW3, and developed the software to run on it. Figure 2 below shows the HW3 board. There are two heterogeneous Arm-based SoCs (System on a Chip) in the middle of this board (green and blue squares) for full redundancy. Inside the SoCs, the CPUs and GPU are Arm-based components, together with a native Tesla-designed NPU. The critical power of both NPUs is a total of 72 TOPS (Trillion Operations Per Second) @ 2 GHz. The inputs from the sensors are on the right of the board, up to 2.5 billion pixels/second. The power supplies are on the left.

This HW3 board installs at the back of the Tesla glove box for all new vehicles and is a replacement for earlier boards. The total power requirement is only 75 Watts. A key performance metric for the HW3 is the number of HD frames per second that can be processed. The HW3 delivers 2,300 fps with NPU hardware running the Tesla software. The same software on HW2.5 is only 110 fps, less than 5% of the HW3 throughput. This rate is insufficient for more advanced ASD features.

For Tesla, this enterprise Matrix workload is core to its mission to produce driving and safety enhancements, and eventually ASD. Tesla believes the HW3 can achieve full ASD. Wikibon believes that at least ten times the power will be needed, as the number of sensors, types of sensors, and the corresponding amount of data dramatically increase over time. The technologies used in vehicles should
allow for additional increases in performance. Tesla is planning to update its technology with 2-3 times more performance by 2022.

**Data Strategy for Tesla HW3**

The second problem to solve is how to set up an end-to-end data strategy that will create an improvement feedback loop at an affordable cost. Some believe that IT will keep all the data from every car. Some have argued that 5G will whisk away all this data for free.

Of course, this data is essential at the very early stages of development. However, some elementary maths by the reader on the billion pixels/second problem will show the fallacy in this thinking when applied to millions of cars and billions of miles over the whole world. Also, that billion pixels/second will grow into many more billions/second in the future.

Wikibon research has pointed out that the vast majority (~99%) of data created at the Edge will be deleted at the Edge because there is no longer any value remaining. Tesla’s approach underlines this projection. The HW3 keeps the data in SRAM for just 10 minutes. During that time, a tiny proportion of the data is sent over the network.

The data selected is to complete the feedback loop from events such as accidents, near misses, pavement anomalies (e.g., pot-holes), and internal and external compliance data. Also, data can be requested from the entire vehicle fleet by engineers working on solutions for specific and rare environments. Examples are how human drivers deal with driving conditions such as snow, sunsets, or large animals on the road. This capability is an innovative and cost-effective solution to the grave “long-tail” problem discussed in the next section.

**“Long-tail” Problem**

There is a very “long-tail” of ultra-low probability events in the overall learning process for ASD5. Access to this data, the precise data about the vehicle components, plus the reaction of human drivers not using or overriding automation from across a fleet of vehicles is essential for developing recognition of rare events and training the ability for correct action.

The Tesla “long-tail” solution allows identification of rare events and the development engineers to receive clip-logged data about them. If Tesla can increase the number of cars on the road from 1 million to many millions, and the miles tracked over the years to billions, the end-to-end data capture and improvement system should deliver exponentially quicker improvements to the overall quality of the ASD system. The end-to-end data architects had to understand and build-in this end-to-end capability at the initial design stage.

This elegant end-to-end data architecture helps solve the long-tail problem and significantly reduces the time and cost to develop full ASD.

MobileEye and Waymo are two other companies that are developing different Matrix technologies and are building up a fleet of vehicles.
**Tesla Conclusions**

The Tesla approach has many positive attributes. These include clear ownership and access to all the data from all Tesla cars, precise end-to-end software and data architecture, and an ability to focus data capture on events that matter. Tesla has designed low-latency electric vehicles with independent electrical engines/brakes integrated and with precise knowledge of the vehicle’s design and internal data. These can react much faster, much more effectively, and much more safely than traditional ICE vehicles. If Tesla can link this internal data to a reliable plan generator, the result is probably an order of magnitude safer vehicles, with much better road utilization. If Tesla can execute and increase its fleet of cars on the road to millions and the miles driven to billions, it can create end-to-end data feedback and improvement loops unique in the transport industry and potentially save millions of lives worldwide.

If successful, Tesla can also use this data architecture for many adjacent business opportunities, including high-level feedback on individual driver safety as an input into the cost of a Tesla-branded insurance policy. Other possibilities include maintenance, leasing, reporting pot-holes, and equipment malfunctions to towns and cities, infotainment, shipping, and journey end-points.

On the negative side, the Tesla case illustrates some of the challenges inherent in developing an early Matrix workload solution, including the level of investment required to create and maintain unique software and hardware. If competitors such as Mobileye and Waymo become volume suppliers, Tesla’s technical overheads may lead to slower adoption of innovation from others in the industry. Other volume suppliers could make it harder for Tesla to convince compliance regulators that the Tesla “long-tail” data solution will work as well as alternatives.

**Bottom Line:** there are a lot of “big ifs” for Tesla and incumbents - and very high stakes for all.

**Other Examples of Edge Disruption**

Whatever the outcome of Tesla’s attempt to disrupt the car industry, Wikibon believes that this type of end-to-end architectural thinking about Matrix applications and data will drive digital innovation and differentiation in many other sectors.

There are many real-time use cases at the business Edge, including robotics, optimization of supply chains, monitoring of wind-farms, automation for civilian and military aircraft, automated warehouses, automated stores, and numerous other examples from every industry. Also, many cross-industry applications, such as advanced analytics, will profit from Matrix application techniques.

Many of these workloads will include at least some primary life-death decision components. Automation of a warehouse can bring potentially fatal interaction between large machines and warehouse staff or external deliverers. Automation of manufacturing with robots brings fast-moving tools near operators with potentially deadly consequences. Automated safety systems on oil-rigs must include the safety of rig workers who cannot escape easily. All these risks make internal and external compliance particularly crucial for developers.

Amazon, for example, is developing this type of technology to automate retail in Amazon Go stores completely. Every enterprise should evaluate its current processes, determine how much more efficient an end-to-end Matrix workload architecture could be and learn how to create end-to-end data architecture to support these workloads.

**Bottom Line:** Wikibon believes that mastery of real-time Matrix workloads and end-to-end data architectures are essential in creating digital differentiation in many industries. Self-disruption is the best form of protection for enterprises against external disruption threats.
Case Study II - Systems of Intelligence & Real-time Pricing

Introduction

So, how should enterprises think about incorporating Matrix technologies into their workload portfolio? Enterprises who successfully implement Matrix workloads with an end-to-end data architecture can address three strategically significant opportunities.

1. They can expect to generate an order of magnitude more additional application value, compared to organizations who continue with the current serialized transaction systems.
2. They can create protection against external disruption from low-cost alternatives, and expand the options available to identify and protect against external disruption.
3. They can expand into adjacent industry areas and be a potential source of disruption.

The challenge is to migrate from current systems of record and data warehouse systems to “Systems of Intelligence.” Wikibon has discussed and defined Systems of Intelligence in previous research. In a nutshell, Systems of Intelligence are systems of record integrated with Matrix workloads such as real-time advanced business analytics and AI.

We will take one example to estimate potential benefits. Currently, most enterprises address pricing changes with a series of steps by professionals. They eventually agree on the new pricing over weeks (or months), and then IT implements these changes at a convenient, quiet time. A section below entitled “The Business Case for Systems of Intelligence and Real-time Pricing” discusses the potential benefits of a real-time Matrix workload, which enables automation of price adjustments compared to the current slower and more labor-intensive manual system.

Traditional Transaction Systems of Record vs. Matrix Workloads

Currently, the systems of record can rarely perform all the workflows in a single transaction. There are multiple serialized transactions, and complex workflows to manage them.

Table 1 below shows the differences between traditional enterprise transactional workloads and updated systems of record integrated with enterprise Matrix workloads. These updates can create a unique solution that can radically simplify workflows and can provide very high levels of automation.

In these Matrix workloads, Table 1 shows the users are likely to include machines and other applications, the control is through the work-flow and APIs, and the data is distributed, highly parallel, and in very high volume. Matrix workloads are most likely to be executed where the raw information is created, which will often be at a business Edge. There will also be Matrix workloads that work best centrally.

The focus of Matrix workloads is to perform many transactions in parallel in real-time or near real-time and will mean much more extensive and distributed databases. The compute model will be cloud-first, hyper-converged, with (in the future) heterogeneous compute architectures (see discussion above entitled “Heterogeneous Compute Architecture”).

Table 1 indicates that enterprise Matrix workloads will drive significant changes in workload hardware and software architecture, as well as data performance and data management improvements.
Table 1: Enterprise Matrix Workload vs. Traditional Transactional Workload Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Traditional Workloads e.g. Systems of Record</th>
<th>Matrix Workloads Systems of Intelligence, Real-time Pricing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Business Problem</strong></td>
<td>Common</td>
<td>Unique</td>
</tr>
<tr>
<td><strong>Users</strong></td>
<td>People</td>
<td>Machines</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>Process</td>
<td>Flow</td>
</tr>
<tr>
<td><strong>Data</strong></td>
<td>Centralized, Serial</td>
<td>Distributed, Parallel, Very high volume</td>
</tr>
<tr>
<td><strong>Facility</strong></td>
<td>Data Center</td>
<td>Edge &amp; Datacenter (where the data is created)</td>
</tr>
<tr>
<td><strong>Time</strong></td>
<td>Serialized Transactions</td>
<td>Real-time or Very-near Real-time</td>
</tr>
<tr>
<td><strong>Database</strong></td>
<td>Heavy access, Small DB Size</td>
<td>Heavy access, Distributed, Very large DB sizes</td>
</tr>
<tr>
<td><strong>Compute Model</strong></td>
<td>Traditional</td>
<td>Cloud-first, Hyperconverged, Heterogeneous</td>
</tr>
</tbody>
</table>

Source: © Wikibon 2020

**Bottom Line:** the benefit of Matrix workloads is a simplification of processes and much higher levels of automation.

**Avoiding Conversions of Systems of Record**

Some vendors and professionals argue that enterprises with current Systems of Record and analytic systems should convert to a “modern” platform. Small companies may be able to start again. The most trivial applications may be easy to rewrite. However, the conversion of large scale systems of record and analytic workloads will bring years of delay and a very high risk of failure. Wikibon strongly believes that building on the application value and data value of existing systems of record is quicker, less expensive, and far less risky. Wikibon recommends that CXOs reject any digital transformation & differentiation strategy that requires the conversion of existing systems of record.

**The Business Case for Systems of Intelligence and Real-time Re-pricing**

Industries such as airlines and hotels use frequent repricing to optimize revenue. Some airlines, like Norwegian, use real-time dynamic pricing.

An excellent example of this real-time pricing in another industry is the dynamic pricing of ride-sharing, introduced by Uber. If there is a low supply of drivers, and prices remain fixed, customer wait-time for rides increases. Customer satisfaction decreases significantly. When Uber raises rates in real-time if demand is higher than driver supply, rising prices reduce need, and availability improves. Service levels remained reasonable, and customer satisfaction is significantly improved because the time to pick up and reliability of arrival at the destination is much more critical for those who choose to pay more. An analysis of the total Uber system shows that the overall benefit of real-time pricing is about 3.5% of gross revenue.

Another extension of automated pricing might be to individualize prices according to the ease of doing business with a customer. If a customer has a higher probability of returning goods or of making
unfavorable social media comments, it might make sense to reduce or eliminate discounts given to these customers.

This real-time pricing example requires access to large amounts of information on all aspects of supply and demand, sophisticated models that analyze this data in real-time, and the ability to apply those changes to the existing systems of record. The potential business benefits are improved revenues, customer satisfaction, and lower costs.

A starting point for calculating the enterprise benefit of real-time pricing would be an increase of 3.5% in gross revenue (from the Uber research above). If all the data is available, and the pricing can be automated, a reasonable estimate might be a factor of 5 reductions in the workforce required to operate pricing.

**Bottom line:** This case study does not conclude that all enterprises should develop real-time pricing as a priority. This case study is just one possible Matrix Inference workload project. However, just starting the process of looking at where Matrix workloads that could increase revenue by a few percents, might improve customer satisfaction, and could reduce operational costs by a factor of 5 will drive innovation. At best, this provides sustainable differentiation and at least ensures competitiveness.

**Enterprise Matrix Workload Implementation Strategies**

Enterprises have a wide variety of software and data assets deployed to assist operational and planning processes, with significant investments across the enterprise in multiple locations and platforms. Wikibon believes that a strategy that attempts to put all the data on one platform in one place will fail for almost all enterprises – most notably, large enterprises. As discussed above, the cost of conversion is crippling and will take many years to achieve.

Instead, Wikibon recommends that enterprises focus on establishing an effective end-to-end data strategy, starting with where data is created today and where it will reside in the future. Enterprises should then implement efficient multi, hybrid, and adjacent cloud strategies to ensure:

- The right data is in the right place at the right time.
- Pushing compute to data minimizes movement and latency.
- The responsibility for deployment and maintenance of infrastructure services is moved to software and hardware vendors and cloud providers, and not by enterprise IT.
- RPA (Robotic Process Automation) and other techniques improve existing processes.

At the same time, enterprises should determine the critical parts of the business where real-time Matrix workloads can make a difference, and be experimenting earnestly to design and develop solutions that will enable radical simplification and automation. Executive management should monitor these projects closely to improve their understanding of the risk of being disrupted and their ability to detect potential disruptors earlier.

**Bottom Line:** Enterprise should identify potential Matrix workloads and develop business cases. Enterprises should focus speeding up systems of record, and plan the migration path to Systems of Intelligence. Robotic process automation (RPA), is another method to streamline operations. Systems of Intelligence can become a foundation for implementing sustainable strategic digital differentiation, dramatically improving business innovation cycle-times. The potential benefits for enterprises will range from ensuring survival to overwhelming long-term business value.

**Matrix Workload Projections**
Key Drivers & Constraints for Matrix Workloads

Figure 3 below is the Wikibon compute-only revenue projections for the emerging Matrix enterprise workloads, shown in grey. The growth is constrained in the first four years by significant constraints in software tools, hardware availability, and skills availability. In particular, it is limited by the lack of enterprise heterogeneous compute systems.

Figure 3 in orange shows the traditional enterprise workloads compute-only revenue projection. Over the second part of the decade, there is significant movement of spend from traditional workloads to Matrix workloads.

Table 3 in Footnotes2 below lists more details on the CAGRs (Compound Annual Growth Rates).

Bottom Line: Wikibon projects that enterprise Matrix workloads compute spend will snowball and become 42% of enterprise compute-spend by the end of the decade.

![Matrix Compute Workloads grow to $48 B (42%) over Next Decade](chart)

**Figure 3: Enterprise Matrix Workload vs. Traditional Workload Projections**

Source: © Wikibon 2020

Conclusions & Recommendations

Case study conclusions

As a disruptor, Tesla has to do things very differently than the current manufacturers. Tesla also has to successfully deal with other very important factors, including manufacturing expertise, the cost of batteries, power management, the availability of chargers, and charging times. These are well beyond the scope of this research.
Wikibon would emphasize two key components of their data processing strategy – the investment in Matrix processing technologies, and the integrated end-to-end data strategy. These are difficult to implement for traditional auto vendors who assemble vehicles from component suppliers and are experts in global supply chains. We suspect that many brands will fail over the next decade.

The Systems of Intelligence case study shows that significant progress can be made to improving the performance of existing systems of record, and start on the process of using existing and new data to provide the advanced real-time analytics to help simplify and automate processes. Wikibon discusses the potential benefit of a specific Matrix workload, automated pricing. Wikibon concludes that the likely estimate of the benefit of implementing real-time pricing updates could be an increase of 3.5% in gross revenue, and a reduction by a factor of five in the operational costs to perform pricing.

**Getting Started on Matrix workloads**

Real-time Matrix workloads will be a significant contributor to digital innovation and differentiation over the next decade. Wikibon believes that a strategy that attempts to put all the data on one platform in one place will fail, as the cost of conversion is crippling and will take many, many years to achieve. Techniques such as RPA (Robotic Process Automation) can automate serialized transaction environments, and make the implementation of real-time Matrix workloads quicker to develop and implement. However, current RPA techniques are not sufficient on their own. Real-time Matrix workloads will make a much more significant impact on systemic simplicity and automation.

Wikibon recommends that the enterprises should focus on establishing an effective end-to-end data strategy, starting with understand the places and processes creating data today and in the future. Enterprises should implement an efficient multi-cloud, hybrid cloud, and adjacent cloud strategies to ensure the right data is in the right place and minimize the transport of data by placing compute with the data. IT should outsource the responsibility for deployment and maintenance of infrastructure services to software and hardware vendors and cloud providers.

Wikibon also recommends that IT should upgrade current systems of record and possibly outsourced some components to improve performance and reduce operational costs. The tier-1 mission-critical databases are often a bottleneck. The only tier 1 large-scale mission-critical databases are IBM DB2, Microsoft SQL Server, and Oracle. All provide cloud options, either on-premises or in their own clouds. The vendor is responsible for performance and reliability and operational upgrades for the complete database stack, from microcode on servers to database upgrades.

It is now possible to run the tier-1 database separate from the operational systems that run the applications that call the tier-1 databases. Both components are in adjacent clouds running in the same datacenter and connected by low-latency direct communications between the two clouds.

For example, Microsoft Azure and VMware offer the ability to run the applications on their cloud, with a very low-latency connection to Oracle adjacent databases also running in the same cloud datacenter. Microsoft and Oracle engineers have worked to provide the connectivity between the two clouds. This allows the flexibility of the database vendor being responsible for using Matrix technologies to speed up the database, and the other clouds to use heterogeneous architectures to accelerate the Matrix inference workloads.

The importance of making such changes is to speed up the current systems of record and allow time for the Matrix workloads to run in parallel. At the same time, enterprises should determine the key parts of the business where real-time Matrix workloads can make a difference, and be experimenting in earnest to design and develop solutions that will enable radical business process simplification and automation. Senior executives should be keeping a wary eye for potential disruptors to their enterprise.
Action Items

The Tesla case study shows the power of an end-to-end data strategy, the investments and commitments necessary to achieve it, and the threat to incumbent auto companies. The case study also shows the power of executing matrix workloads to develop ASD and build continuous and potentially exponential improvement processes.

Wikibon strongly recommends that senior enterprise executives develop an end-to-end data strategy, and develop a deep understanding of Matrix workloads. It is important that the approach focuses on real-time data flows, and not on historical data, which is likely to be of much lower value.

This exercise will allow senior management to understand where external innovation threats are most likely to come from, and identify early where Matrix workloads could be used by external disruptors. This allows senior management to understand the level and nature of any disruptive threat, and to either acquire, partner, or bring forward internal projects to ensure the enterprise remains competitive.

In addition, the development of an end-to-end data strategy can help to avoid losing control of access to key data. This can happen if part of an organization does not understand the loss to the company as a whole if it gives away data to a supplier or potential competitor.

Upcoming Matrix Workload Research

Scope

The next Wikibon research report on Matrix workloads will be entitled “Heterogeneous Compute Accelerates Matrix Workloads”. The research will look at the changes in compute architecture required to enable Matrix workloads to operate in real-time, with 10 to 1,000 times more data than today’s transactional systems.

Premise

The simple premise of this research note will be that heterogeneous compute architecture can increase the performance of Matrix workloads by a factor a 50, the price-performance by a factor of 200, with 200 times better power & cooling costs. These improvements dramatically increase the range of matrix application inferences that can be real-time and the quantity and quality of the dataflows.

Publication Date

The publication date is scheduled for the middle of March.

Feedback

Wikibon encourages feedback on this research on the subject of Matrix workloads. You can contact the author by email at David.Floyer@Wikibon.org, by Twitter at @dfloyer, and also on LinkedIn.

Footnotes:

Footnote 1: Wikibon Matrix Workload Definition

Table 2 below gives a more formal Wikibon definition of Matrix workloads compared with traditional workloads.
Table 2: Wikibon Matrix Workloads Definition

Source: © Wikibon 2020

Footnote 2: Matrix Workload vs. Traditional Workload Compute Revenue Projections and CAGR Table

Table 3 below gives the worldwide compute revenue for Matrix workloads and traditional workloads 2019 - 2030. The CAGRs 2019 - 2030, CAGR 2019-2025, and CAGR 2025-2030 are also shown in the table. As expected, the CAGR growth of Matrix workloads is the fastest 2019-2025 (CAGR 42%), but the year-on-year actual revenue growth continues to increase all the way through to 2030 ($11B growth 2029-2030).

The main business driver for this growth is the very high business returns on Matrix workloads. The main inhibitors to growth are a development learning curve and the availability of hardware and software platforms.

Table 3: Matrix Workload vs. Traditional Workload Compute Revenue Projections and CAGR Table

Source: © Wikibon 2020
David Floyer spent more than 20 years at IBM, holding positions in research, sales, marketing, systems analysis and running IT operations for IBM France. He worked directly with IBM’s largest European customers, including BMW, Credit Suisse, Deutsche Bank and Lloyd’s Bank. Floyer was a Research Vice President at International Data Corporation (IDC) and is a recognized expert in IT strategy, economic value justification, systems architecture, performance, clustering and systems software.