Part 1 - Current and future Persistent Memory related enhancements on IBM Power Systems™ for SAP HANA

By Asim Mustafa Khan – Senior Offering Manager SAP HANA on POWER

It shouldn't surprise anyone that IBM's strategy for existing and upcoming memory technologies is anchored on embracing the open ecosystem that gives clients choices to solve real problems. When IBM and SAP jointly announced support for SAP HANA on IBM Power Systems 4 years ago, IBM did not simply follow what every other hardware vendor was doing. Instead, we analyzed the problems that clients had and created solutions that solved them with virtualization and flexibility, resiliency and lots of performance to top it all off. The market reaction after these years shows thousands of clients embracing IBM Power for SAP HANA (most who left their antiquated x86 appliances behind) and we have experienced unparalleled growth in the Enterprise Linux market.

Our approach is very similar when it comes to persistent memory. Instead of embracing a single persistent memory technology and vendor – as the x86 competition seems to be doing – we will execute on a multi-step approach that enables us to work with multiple memory technology vendors and have a more comprehensive range of solutions that clients will be able to deploy, in some cases, without even having to purchase new hardware at all.

Client requirements and pain points

Thousands of clients have shared with us the pain points they suffer as they attempt to tackle the costs of managing their HANA on x86 deployments. In general, the typical statement is:

'frequent planned or unplanned outages force lots of wasted time waiting for systems to shut down, be brought back up, patched and all data reloaded'

All x86 vendors have answered in the same way - pretending like downtime is normal and promising memory that sacrifices performance in order to provide persistence and faster load times. At IBM, we have gone deeper and realized that:

- Clients demand choice, and all x86 vendors who are embracing a single persistence technology from Intel are, in essence, providing the same single vendor proprietary technology.
- Our current SAP HANA on IBM Power clients do not have nearly as many server outages, but even they would like to be able to load faster after patching the stack.
- Clients (on x86 or Power) do not want to tradeoff performance for persistence.
- Clients want to add the capability of persisting memory without having to rip and replace or purchase expensive add-ons to hardware they just bought.

IBM Power Systems vision for Persistent Memory

With these client pain points in mind our vision is one that starts with:

- **What is possible today [LINK].** Embracing the latest technology advancements that SAP has introduced and that run much better and faster on IBM Power Systems, such as 'tmpfs' and Native Storage Extensions.
- **What is coming later this year [LINK].** PowerVM with PMEM will be the first announcement planned for the end of 2019 that will help address all the above business requirements. PowerVM with PMEM (Persistent MEMory) is an enhancement in our Virtualization platform that will create persistent memory volumes using the existing DRAM technology that our clients already own. By maintaining data persistence across application and partition restarts, it will allow our clients to leverage fast restart of a workload using persistent memory for the vast majority of their planned maintenance and unplanned outages without compromising the performance of HANA during normal use. This capability will be available without changes to existing applications.
- **What will come in the next 2-3 years [LINK].** There are many new memory technologies planned to hit the market that will likely disrupt the industry, providing persistence characteristics with a range of performance and cost that will lower the hardware costs without slowing down in-memory workloads like SAP HANA. These technologies will come from a multitude of memory partners, not a single one, and IBM will be right in the middle of it leveraging industry standards.

Please read the rest of this document to go deeper into each of these topics and get a better understanding of our vision and how IBM and SAP execute on that vision to create choice and allow our clients the flexibility they need to unleash the Power of SAP HANA.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion. Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

Part 2 - SAP Native Storage Extension and Fast Restart – what is possible today

By Wolfgang Reichert – IBM Distinguished Engineer and CTO for SAP on IBM Systems

With the continuous growth of business data, customers look for new technologies to lower total cost of ownership (TCO). At the same time the business demands higher level of service. SAP and IBM released new features and technologies to address such customer needs.

1. TCO reduction

A significant part of the server cost is related to the amount of memory (DRAM). The more data are kept in memory the higher the price. Depending on the chosen license model, the in-memory footprint of HANA may also affect the SAP software license cost. Other data tiering options have been already available though they are only applicable to rarely used data. Furthermore, these tiering options require a second LPAR and connectivity to the main HANA instance causing higher cost and complex operations.

The Native Storage Extension – a new SAP HANA feature – allows to reduce the in-memory size of the database by utilizing a buffer cache and keeping less frequently used data on disk. This is a build-in functionality in HANA that just needs to be switched on.

2. Reduce planned downtime

Planned maintenance – e.g. for SAP HANA or Linux software updates – requires restart of the database. For multi-terabyte databases the restart and data load time contributes significantly to the overall downtime. The larger the database the longer it takes.

The new SAP HANA Fast Restart Option dramatically reduce the restart time of the database.

In addition, IBM is currently implementing PowerVM with PMEM (virtual persistent memory) to further extend the scope of a fast restart.

3. Scale up to larger database sizes

Until now, growing the database required an increase of HANA system memory. If the maximum of supported HANA memory size per node is exceeded a scale-out configuration is required.

The Native Storage Extension allows to leave less frequently used data on disk, e.g. aged data. The data on disk do not count for the in-memory size, in fact, this allows to grow the database significantly over past limits.

Combined with low-latency storage attachment the overall HANA performance is preserved.

SAP HANA Fast Restart Option

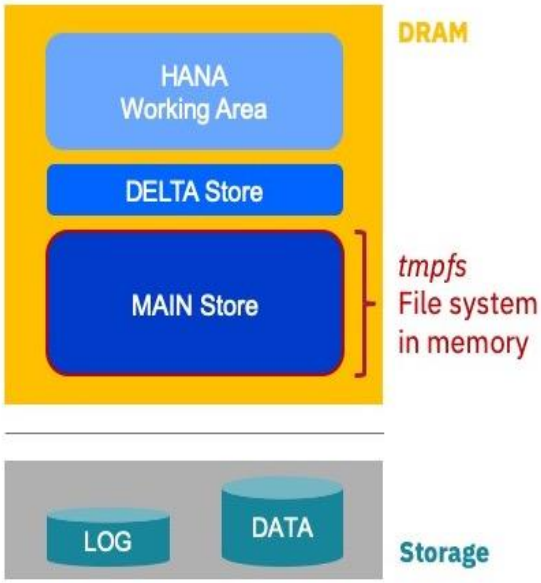
Whenever HANA is restarted all data in memory are cleared and need to be reloaded from storage. The larger the database the longer it takes.

Actually isn't this wasted time? Why not preserving the data in memory?

Linux offers a temporary filesystem *tmpfs* that resides in memory and survives application lifecycle. With this new feature, HANA moves large parts of the memory content to *tmpfs*. At startup, HANA checks if retained memory content is consistent and can be used instead of loading from storage.

No need for additional memory, no need for special hardware. And there is absolutely no impact on performance or latency at runtime because *tmpfs* is in DRAM. (The file system just serves an administrative overlay structure to identify memory pages and map them as tables, columns, etc. at HANA startup. It is alike the layout on storage.)

What benefits can be seen measuring HANA shutdown and restart time on IBM Power? For this purpose, we loaded 8 TB of data. It represents a 16.TB HANA system considering that the working area is as big as the data area.



In the standard setup, HANA shutdown takes about 6 minutes. HANA startup including data load takes more than 40 minutes. (Note: HANA is up and responsive sooner, but we check for the log entry indicating that all data have been loaded.)

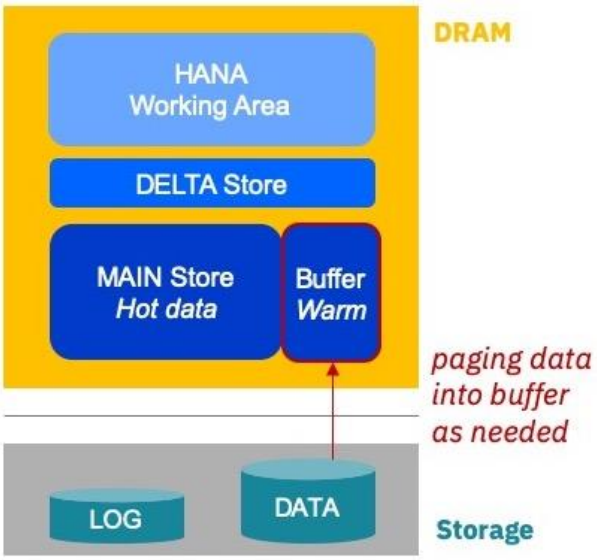
With the Fast Restart Option, the startup time is drastically reduced. HANA is fully available within less than 4 minutes. Also shutdown is faster as can be seen in Figure 2.

Noticeably, the mapping and un-mapping of very large amount of memory into the HANA address space benefits from the 64K page size on IBM Power Systems, providing a unique platform advantage.

SAP HANA Native Storage Extension

In prior HANA releases data growth and data tiering have been addressed by scale-out configurations and/or data archiving.

With HANA 2 SPS04 a new feature has been introduced that allows to keep less frequently used data ("warm data") on storage instead of loading all data into memory. This feature enables a substantial increase in SAP HANA data capacity without impacting performance for high-data volumes. Customers will be able to expand SAP HANA database capacity with warm data on disk up to about four times the size of hot data in memory.



The Native Storage Extension (NSE) is part of the HANA data tiering concept, that is the assignment of data to different storage and processing tiers based on various characteristics of the data (hot, warm, and cold data for these tiering systems).

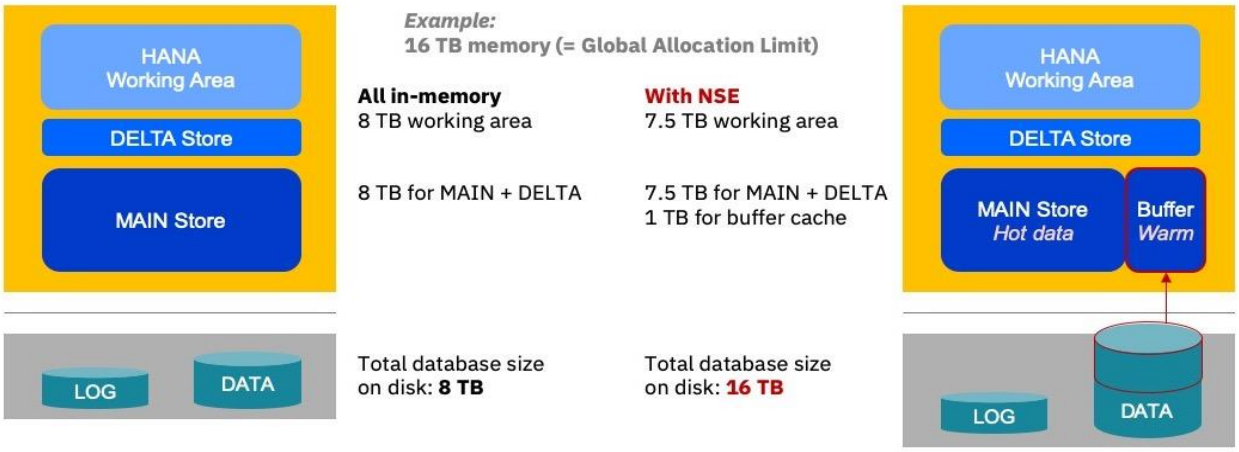
NSE is deeply integrated with the HANA database. It implements a buffer cache for HANA column store tables. Warm data mainly reside on disk and loaded into memory only as required for query processing.

Currently SAP has established the following sizing rules. As indicated, the rules will be relaxed based on customer experience with this new feature.

- HANA system must be scale up (first release restriction)
- May add as much warm storage as desired – up to 1:4 ratio of HANA hot data in memory to warm data on disk
- NSE disk store should be no larger than 10TB (first release restriction)
- Divide volume of warm data by 8 – this is size of memory buffer cache required to manage warm data on disk
- Work area should be same size as hot data in memory

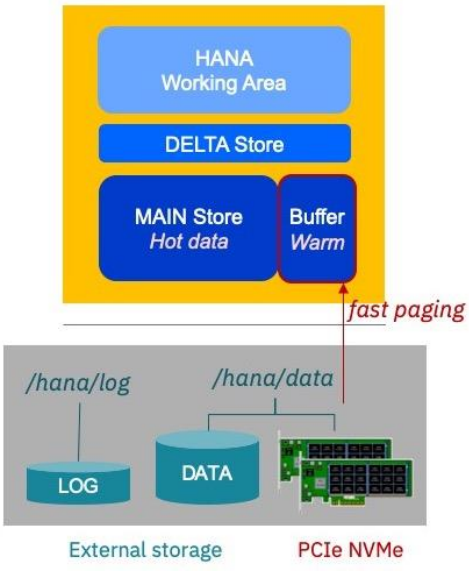
Based on above rules, the following advantages are expected:

1. The maximum of supported HANA database size that can be managed by a scale-up system is roughly doubled – without changing the server hardware or software licensing. In the future, even more data on disk may be supported.
2. Lab measurements with HANA have shown the following savings: Given a mixed workload and data, 50% of column tables can be marked as warm data without noticeable performance degradation. Thereby the global allocation limit (GAL) can be reduced by 38% – resulting in smaller HANA system size.



Accelerated storage attachment

A low-latency storage attachment allows faster paging of warm data as well as quick startup of HANA. For POWER9 systems the new PCIe3 x8 NVMe Flash Adapter (add-in card) can be used. NVMe is a high performance software protocol that can read/write flash memory. Compared to a SAS or SATA SSD technology, the NVMe Flash adapter provides more read/write input/output operations per second and larger throughput at lower latency.



The PCIe NVMe cards are available with capacities of 3.2 TB and 6.4 TB. The POWER9 servers are equipped with several PCIe slots; the exact number depends on the model.

The PCIe NVMe storage can be configured as file system mirror for the persistency on external storage devices, as shown in the figure. Hence backup, high-availability configurations and lifecycle management remain unchanged.

The NVMe flash storage can also be used as local disks, assuming that appropriate data replication and/or backup procedures are in place.

References

SAP HANA Administration Guide for SAP HANA Platform, Version 2.0 SPS 04

- SAP HANA Fast Restart Option

  (https://help.sap.com/viewer/6b94445c94ae495c83a19646e7c3fd56/2.0.04/enUS/ce158d28135147f099b761f8b1ee43fc.html)

- SAP HANA Native Storage Extension

  (https://help.sap.com/viewer/6b94445c94ae495c83a19646e7c3fd56/2.0.04/enUS/4efaa94f8057425c8c7021da6fc2ddf5.html)

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion. Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

Part 3 - PowerVM with PMEM (Persistent Memory)

By Asim Mustafa Khan – Senior Offering Manager SAP HANA on POWER

PowerVM with PMEM is planned before the end of 2019 and will address the typical business requirements that clients have shared with us:

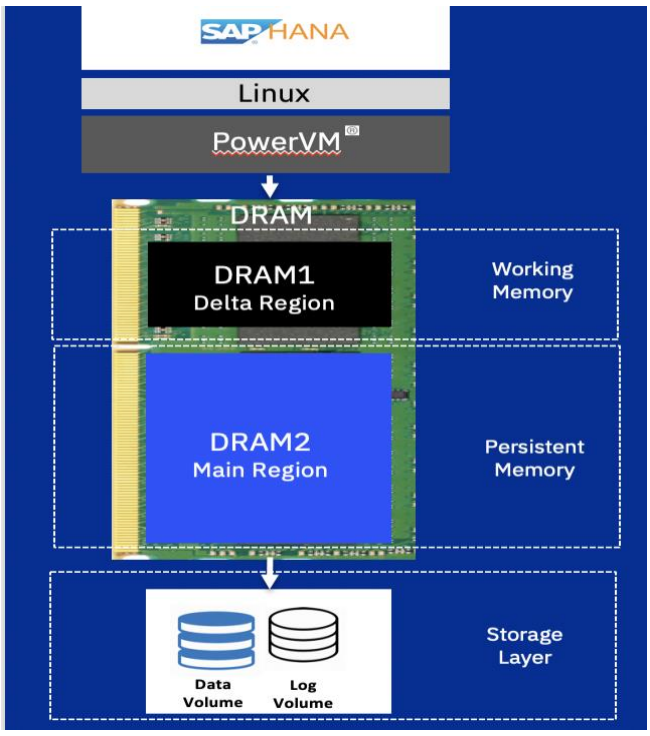
- Clients demand choice, and all x86 vendors who are embracing a single persistence technology from Intel are, in essence, the same proprietary technology.
- Our current SAP HANA on IBM Power clients do not have nearly as many server outages, but even they would like to be able to load faster after patching the stack.
- Clients (on x86 or Power) do not want to tradeoff performance for persistence.
- Clients want to add the capability of persisting memory without having to rip and replace or purchase expensive add-ons to hardware they just bought.

PowerVM with PMEM (Persistent MEMory) is an enhancement in our Virtualization platform that will create persistent memory volumes using existing DRAM, maintaining data persistence across application, operating systems and partition restarts. This capability requires no changes to existing applications.

Persistence without compromising performance, virtualization and TCO

Since this feature is based on existing DRAM technologies, it has the same performance characteristics clients already experience today, providing clients with the peace of mind of using technology that already meets their performance requirements.

The table below indicates some of the typical comparisons that clients may want to use when comparing persistent solutions, in particular, I'd like to highlight that IBM PowerVM with PMEM will shrink downtime windows significantly; even more than other technologies by drastically reducing shutdown as well as start up due to technical capabilities only available on IBM Power Systems.



Scenario	Intel® Optane™ Technology	PowerVM with Virtual PMEM (IBM POWER9)
Speed up HANA restart	Fast	Fastest
Speed up HANA service restart	Fast	Fastest
Speed up HANA upgrade/service patch	Fast	Fastest
Available on existing hardware w/o extra cost	No, Cascade Lake Processor based systems only GA 04/2019	Yes (Power9 sold since 2018)
Virtualization is supported	No	Yes
Preserves runtime performance/latency	No	Yes
Improves Shutdown time of SAP HANA environment	No (degrades it)	Yes

Table 1

You may be wondering why we think that persistence at the LPAR level is the best solution. We could have aimed to match what the x86 competition was planning, but the loud feedback we received from our existing HANA on Power clients was:*'why try to solve a problem you don't have?'*

Clients who had lots of experience with x86 and Power for SAP HANA told us that, when they switched to using Power, hardware related outages became very rare (unlike when HANA is running on x86 based servers). When using Power, the vast majority (if not all) of the outages that the clients have are for operational needs like OS or application related upgrades and patching. And they needed to reduce this maintenance window to improve availability of the environment to their businesses without the fear of degrading everyday performance. Our PowerVM with PMEM solution meets and infact exceeds all these requirements.

Unlike x86 vendors, we will not force our existing SAP HANA on POWER9 clients to rip and replace their recently procured hardware and pay to have this feature. It will be available to all clients running POWER9 based servers with a simple software update at no additional cost!!

Since this feature will be supported with Virtualization on Power Systems clients will continue to have the same flexible scalability and granularity they already enjoy along with dynamic and flexible CPU and Memory resource assignments to each VM on the system. If the clients want to

deploy an x86 persistent solution, it will have to be baremetal as virtualization is not yet supported.

We believe clients running SAP HANA on POWER will continue to have significantly better TCO than our competition without the trade offs that are forced by their proprietary persistent memory technologies.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion. Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

Part 4 - Future Memory enhancements on IBM Power Systems

By Asim Mustafa Khan – Senior Offering Manager SAP HANA on POWER

Historically, a server has been able to use one generation of DRAM technology (i.e. DDR4 DRAM) and one or two storage technologies (e.g. Flash SSD and Hard Disk Drives). Today and over the next several years, there will be many memory technologies that will have differences in performance characteristics and cost. All will have varying use cases based on system applications.

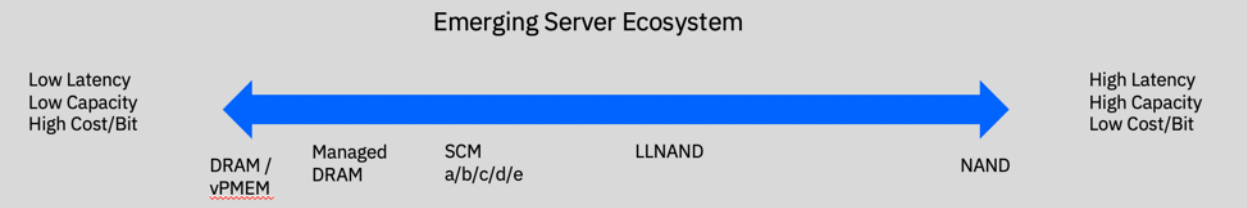


Figure 1: Emerging Memory Technologies

There are several emerging memory related technologies that are and will become available in the industry (see Fig. 1) addressing different needs and forcing tradeoffs. For example:

- Using the current DRAM with low capacity DIMMs, clients can get highest data throughput performance due to low latency, but at higher cost than SCM or NAND technologies.
- Substituting DRAM by SCM will help reduce costs but will increase the latency.
- Using NAND technology for storing volume data will help reduce the cost per GB even further with the disadvantage of a magnitudes higher latency compared to DRAM or SCM technologies.

As the industry is beginning to exploit storage class memory, IBM is developing a high performance **Hybrid Memory Subsystem (HMS)** offering. The HMS adapter will leverage off-the-shelf technologies to deliver high capacity, low cost, byte-addressable persistent memory and in addition exploiting the several SCM technologies that are available in the system. This technology will address the low cost based client requirements and is built upon OpenCAPI based interface to provide better throughput. The main benefits of HMS technology are:

1. It does not displace main memory DIMMs, thus enabling IBM Power Systems to achieve higher system memory capacity, on top of main memory DRAM
2. It does not require DRAM DIMMs on the same memory channel to achieve performance
3. It is compliant with PMDK PMEM/DAX interfaces, offering IBM customers plug-and-play compatibility with applications built for common persistent memory offerings.

IBM's roadmap for Storage Class Memory (SCM) begins with LLNAND and extends to future SCM technologies. These technologies are not mutually exclusive and address different cost/performance targets. That's why IBM Power Systems customers will continue to get the flexibility to choose the most appropriate memory technology for their unique workloads.

In addition, the POWER9 processor available today already takes some of the pain of reloading a large database into memory leveraging its robust I/O subsystem. POWER9 systems have PCIe Gen4 technology that doubles the I/O bandwidth compared to competitive processor technologies that are based on PCIe Gen3, thus resulting in much better performance when using the SAP's newly announced technologies like NSE (Native Storage Extension) that uses the system I/O bandwidth.

It is also important to note, however, that current persistent memory technologies are not a replacement for existing data-at-rest storage. Our clients expect multiple layers of redundancy when it comes to the storage of data-at-rest. That's why the industry leading IO subsystem in POWER systems provides our clients with the advantage of being able to permanently persist and reload data faster than competitive systems after planned and unplanned outages.

As the technology enhancements in memory continuously evolve, IBM is working with several industry leaders in this space and making sure that the right open standards technology (DRAM, PowerVM with PMEM, SCM and LLNAND) is made available to our clients in timely fashion and that we provide more choices than our competition.
We will continue to execute under the premise that clients want choice and we need to deliver 'Unmatched Flexibility' to our clients. Sticking to our principle of focusing on client needs and pain points, and not falling in the mass mentality of trying to solve them like if we were one more x86 vendor. We provide needed choices for a healthy ecosystem.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion. Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.