IBM Power S1012

A 1-socket, half-wide system designed for both edge computing and core business workloads

Highlights

Optimize data center space and run workloads at the edge

Improve IT economics for business-critical workloads

Secure data and insights in and out of AI models

Embedded AI acceleration for faster inferencing

The core applications, data stores, and processes that run your business simply cannot go down, no matter what. With the accelerator of digital adoption, the demands on those are increasing, along with the related security risks. To stay ahead of the curve, your IT systems need to be modernized to meet the challenges of today. This requires an infrastructure platform that efficiently scales to meet new demands, protects your applications and data with pervasive and layered defenses, and enables you to transform data into insights quickly.

The IBM® Power® S1012 is the edge-level server of the IBM Power portfolio. The S1012 is a 1-socket, half-wide, Power10 processor-based system designed for both edge computing and core business workloads. It is available in a 2U rack-mounted or tower chassis form factor and delivers the lowest entry price point in the Power portfolio to run core workloads for small and medium-sized organizations. The Power S1012 provides clients the flexibility to run AI inferencing workloads in remote office and back-office locations outside mainstream data center facilities, and in direct connection to cloud services such as IBM® Power® Virtual Server for backup and disaster recovery.

Benefits

Reduce physical footprint
The Power S1012 features a half-wide design in the rack-mounted form factor. Clients can optimize space in the data center by running two systems side-by-side in the same rack space. With the half-wide design, clients can reduce space allocated for IT infrastructure by up to 75%[1]. This allows support for application modernization projects in a smaller space. Other use cases include development and test environments in the same rack space, remote IT management by utilizing one system as a vHMC next to a production environment, high availability with IBM PowerHA, and running AI inferencing models at the point of data.

Increase performance of applications and AI models
The Power S1012 has up to 3X more performance per core compared to previous generations[2]. Clients can improve performance of business-critical applications while optimizing space in the data center or running workloads in locations that have limited infrastructure or the traditional space needed for servers. By running AI inferencing at the edge close to where critical data is being generated, latency is reduced and performance of models is improved. Power S1012 also has built-in AI acceleration to improve performance of models without the costly expense, maintenance, or noise disturbance associated with GPUs.

Run AI inferencing workloads at the edge
The Power S1012 is designed to enhance remote management capabilities for clients looking to expand applications such as AI inferencing from core to cloud and at the edge. Edge computing can also provide a competitive advantage with real-time insights across industries, with examples that include analyzing customer behavior in retail, monitoring and optimizing production processes in manufacturing, and many more. With the half-wide design, the Power S1012 is ideal for workloads and deployments at the edge in remote office and back-office locations with limited space and infrastructure such as a retail store, hotel, hospital, manufacturing facility, and more. Clients can gain real-time business insights by running AI inferencing models at the point where data is being generated and reduce latency associated with needing to connect back to a centralized data center or cloud environment.

Lower IT infrastructure costs
IBM Power10 processor-based servers can consolidate multiple workloads onto a single system, reducing the number of servers needed and resulting in lower hardware, software, and maintenance costs. Built-in PowerVM enables multiple virtual servers to run on a single physical server, leading to better resource utilization and reduced energy consumption. By upgrading and consolidating environments onto fewer systems, clients can lower electrical, cooling, and costs associated with data center space, along with reducing carbon footprint. With 1, 4, or 8 cores available, small to medium size clients can upgrade infrastructure at a lower price point and take advantage of the security, reliability, availability, and performance of Power10 processor-based servers.

Industry-leading RAS
IBM Power has been ranked the most reliable non-mainframe platform for the past 15 years by ITIC[3]. An ITIC survey of 1,900 C-level executives across 37 industry vertical markets gave IBM Power a 99.9999% or greater availability rating. IBM Power servers feature redundant components, such as power supplies, fans, and disk drives, which ensure that the system remains operational even if one component fails. Power servers also feature advanced diagnostic capabilities, including predictive failure analysis and automatic error reporting, which enable proactive maintenance and reduce system downtime.

75%

With the half-wide design, clients can reduce the space allocated for IT infrastructure by up to 75%.

3X

The IBM Power S1012 has up to 3X more performance per core compared to previous generations.

"IBM Power S1012 will provide the latest capabilities to support AI inferencing where the data itself is generated."

Matt Niessen
President
Equitus

Features

Half-wide design
The IBM Power S1012 is the smallest Power processor-based server ever built and has a half-wide design in the rack-mounted form factor. In the data center, clients can optimize space by utilizing two Power S1012 servers next to each other in a single rack space, allowing for several use cases such as development and test, high availability, remote IT management, and more. Outside of the data center, clients can run workloads at the edge in remote office and back-office locations that have limited space and infrastructure, such as retail stores, hospitals, hotels, manufacturing facilities, and more. This allows applications and AI models to be run where data is being generated, improving performance and availability of these workloads.

1, 4, or 8 processor cores
The IBM Power S1012 has the lowest entry price point in the Power portfolio, ideal for clients looking to upgrade infrastructure affordably to take advantage of the benefits and capabilities of Power10. It is available with 1, 4, or 8 processor cores to improve IT economics and efficiency for business-critical applications. Clients can consolidate databases and business-critical applications onto fewer servers with the improved performance of Power10, reducing costs associated with data center space, electricity, cooling, software licensing, and maintenance. This also reduces carbon footprint for clients that must meet corporate or government-mandated sustainability objectives.

Embedded AI acceleration
Each Power S1012 includes four Matrix Math Accelerators per core to support AI inferencing at the point of data. This built-in AI acceleration allows clients to run models at the point where data is being generated (such as a retail store) without the costly expense, maintenance, and noise disturbance of GPUs. Clients can gain real-time business insights by running these models close to where critical data is being generated and act on these insights to gain a competitive advantage and improve customer, partner, and patient experiences. Running AI models at the edge close to data also reduces latency and improves performance of models by eliminating the need to connect back to a centralized data center or cloud environment.

Transparent memory encryption
Secure the data and insights in and out of the AI models running locally and prevent data leaks. Transparent memory encryption is a security feature available on all Power10 processor-based servers, including the Power S1012, that encrypts data in memory to protect it from unauthorized access. This feature is designed to provide an additional layer of security for sensitive data, without requiring changes to applications or operating systems. Use transparent memory encryption with no additional management setup or impact to performance of workloads and applications. Transparent memory encryption can also help organizations meet compliance requirements, such as those that mandate the protection of sensitive data.

Concurrent maintenance
The IBM Power S1012 has concurrent maintenance features and capabilities to increase RAS and reduce downtime for clients. Concurrent maintenance is a feature of Power processor-based servers that allows for maintenance and upgrades to be performed on a system while it remains online and available to users. This means that IT staff can perform tasks such as firmware updates, hardware replacements, and software patches without requiring a planned outage or downtime. Many Power server components, such as disk drives, power supplies, and fans can be replaced or added while the system is still running, minimizing downtime and reducing the risk of system failure.

Use Cases

IBM i

For small to midsize IBM i clients, the Power S1012 provides a path to the latest Power10 processor-based technology. Available with 1,4, or 8 processor cores to improve IT economics and efficiency for business-critical IBM i applications, clients can improve performance up to 3X versus the Power S814. Moreover, the Power S1012 2U half-wide design can reduce space allocated to a client's IT physical footprint by up to 75% versus the Power S1014 4U rack server and allows for application and IT management convergence with the option to use one server for production side-by-side with the other hosting a virtual hardware management console (vHMC) or to be used for development, test, or high availability.

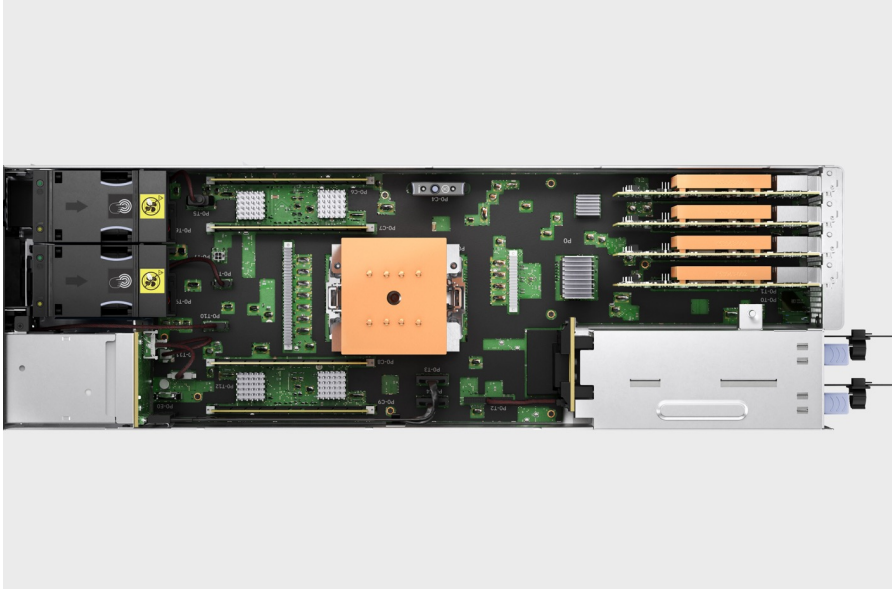
Small databases

Clients running Oracle Database Standard Edition 2 on AIX and other small databases on past generations of Power processor-based servers will see huge benefits of upgrading to the Power S1012. Reduce Oracle licensing and support costs by consolidating database workloads onto fewer servers with Power10. With concurrent maintenance, clients can improve RAS by avoiding system outages and keeping applications online while updates are being made to the hardware. Clients can reduce concerns with data privacy and data loss and support application modernization projects in a smaller space.

IBM Power S1012 Tower

AI inferencing at the edge

By deploying Power S1012 at the edge, clients can run AI inferencing at the point of data, thus eliminating data transfers. This system is perfect for small language models and vision models deployed outside of the data center in locations such as retail stores, hospitals, hotels, manufacturing facilities, and more. Each Power S1012 includes four Matrix Math Accelerators per core to support AI inferencing. This built-in AI acceleration improves the performance of models running at the edge without the cost, maintenance, or noise disturbance of GPUs. To ensure insights remain a competitive advantage and don't fall into the wrong hands, transparent memory encryption with Power10 secures data in and out of AI models running locally to address data leaks. Moreover, with advanced remote management capabilities and Power10 best in class reliability features, IBM Power S1012 allows organizations to efficiently manage and monitor their IT environments remotely to enhance responsiveness and minimize downtime. High-availability features such as redundant hardware and failover mechanisms can help ensure continuous operations, all within a compact physical footprint.



IBM Power S1012 Overhead

Support and Availability

Maintaining high availability throughout the life of systems like IBM Power S1012 is critical. IBM Power Expert Care offers a way of attaching services and support through a tiered approach right away. Clients can receive an optimum level of support for the mission-critical requirements of their IT infrastructure with options ranging from 3 to 5 years of coverage depending on the support tier. Additionally, there are optional committed service levels available, depending on client needs, which can provide further customization and support.

IBM Power S1012	**S1012** **MTM: 9028-21B**
Sockets	1 eSCM
Socket Power Max	240W (rack), 195W (tower)
Module Core Counts	8 (rack only), 3, 1[2]
Core Count (SMT8) Max per system	8
Memory Slots	4 DDR4 ISDIMMs, memory buffers down on planar
Memory capacity	256 GB
Memory bandwidth	102 GB/s
System PCIe slots	2x G5 x8 or G4 x16 direct 1x G5 x8 direct
Drives	4 NVMe U.2 Optional RDX

.

For more information

To learn more about IBM Power S1012, contact your IBM representative or IBM Business Partner, or visit ibm.com/products/power-s1012

1. Based on moving from 4U IBM Power S1014 to a half-wide 2U IBM Power S1012
2. Based on the CPW Benchmark results for 1-core of 29,000 on IBM Power S1012 compared to 9,360 on IBM Power S812
3. https://www.ibm.com/account/reg/us-en/signup?formid=urx-39584

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
May 2024