ScaleMP™
Creating the Power of One

THE vSMP™ ARCHITECTURE AND vSMP FOUNDATION AGGREGATION PLATFORM

WHITE PAPER

BACKGROUND

A BRIEF HISTORY OF HIGH-END COMPUTING SYSTEMS

Manufacturers, designers, life science researchers, financial institutions, federal and military units, geoscientists and artificial intelligence data scientists all demand faster and larger systems for high-performance and computationally intensive applications. Data-driven organizations like banks, retail chains and telecom service providers demand shorter processing time for data, in ever-growing databases.

Since the late 1980s, system vendors were able to accommodate such applications by offering two types of multi-processor systems:

- Shared-memory multi-processor, enabling all system processors to access the entire memory address space. Such shared-memory systems could be either symmetric multi-processors (SMP, where memory "distance" from or latency to- each processor is identical) or non-uniform memory access (NUMA, where memory distance differs based on the accessing processor). In this document, we use the term SMP for shared-memory multi-processor in either SMP or NUMA architecture.
- Distributed systems, such as massively parallel processing (MPP) systems, or clusters, where processing and memory power is distributed between separate computers connected with a high-performance fabric — an approach that mandates the use of special programming techniques to pass messages between the application fragments running on different systems.

SMPs have many advantages: They are easy to deploy, manage and program for; and they can run any workload, including the distributed ones. These advantages are sometimes overshadowed by drawbacks of SMP products, which stem from the proprietary nature of shared-memory systems — first and foremost their price: They use custom chipsets and ASICs to create high-speed memory-coherent backplanes, and sometimes use custom processors or custom operating systems.

Distributed systems deliver more performance per dollar than SMPs, but are more difficult to deploy and manage: The fabrics and the many nodes are not simple to manage, nor are parallel file-systems to support the I/O demands for those configurations. Furthermore, distributed systems can only be used for a subset of the applications, while other applications require SMP resources.

HIGH-END SYSTEMS OF OUR DAYS

Faster commodity processors became available in the 1990s, offering a viable alternative to proprietary processors. This trend accelerated in 2002, when high-performance commodity cluster implementations were adopted, interconnecting commodity servers with commodity, high-speed interconnects. But these solutions were similar to MPP systems, requiring a difficult programming model to allow the application to span across multiple systems.

Customers remain frustrated. Cluster implementations are cost-effective, but complex and limited in applicability; SMPs could be a good choice, but they're expensive.

VIRTUALIZATION FOR HIGH-END COMPUTING

ScaleMP's patented Versatile SMP™ (vSMP) architecture uses off-the-shelf, industry-standard servers and interconnects to create a virtual SMP system with superior capabilities to proprietary SMP systems, while maintaining the commodity clusters' cost structure. vSMP architecture is processor-neutral and agnostic to interconnect technology.

ScaleMP's vSMP Foundation software implements the vSMP architecture, and aggregates up to 128, x86 servers into a single virtual machine, with up to 1,024 processors (as of 2019, more than 30,000 cores) and 2PB of main memory. It supports the latest generations of Intel and AMD processors, both server and desktop variants. vSMP Foundation provides the largest system memory available in the industry, as well as RAS features such as redundant InfiniBand backplane, fast detection and isolation of failed hardware components and partitioning support. As a software-based architecture, vSMP Foundation is inherently software-defined and on-demand: Users can aggregate (part of) the nodes in their cluster into a shared-memory system for a specific workload or project.

> **"ScaleMP offers a great alternative to partners, system integrators and end users looking for a non-proprietary option that would combine the benefits of scale-up and scale-out architectures while preserving their existing investments."**
>
> **Joseph Martins, managing director, Data Mobility Group analyst firm**

ScaleMP's vSMP Foundation enables next-generation, affordable SMP systems for computer manufacturers and end users. Software-based SMPs can be created without investing tens or hundreds of millions of dollars in proprietary R&D, and without losing valuable time to market. Users can create low-cost SMP systems using commodity x86 servers and standard interconnects that deliver the lowest overall total cost of ownership by:

- Running any type of HPC application, providing best-of-breed performance for SMP applications
- Leveraging cluster cost benefits, eliminating the need for custom hardware and components
- Creating SMP on demand using a single management point, thus increasing utilization and lowering costs
- Using the latest generation of processors and interconnects to provide best performance at volume pricing
- Providing selective scaling capabilities so that users can shape the system to fit the workload and only pay for what they use

CHALLENGES OF HARDWARE-ONLY SMP SYSTEMS

SMP SYSTEM COST

SMP systems with more than eight processors cannot be made using off-the-shelf chipsets, and in some cases, custom chipset are required for even smaller systems. They typically require annual R&D investments of tens to hundreds of millions of dollars, as well as substantial development time (measured in years) to bring the solution to market. To recuperate the costs, vendors charge a high premium to a relatively small number of end customers.

PROPRIETARY SYSTEMS

Proprietary technology based on custom hardware and components is expensive. Some SMP systems with four processors or more utilize non-x86 processors — a major contributor to these systems' high price tag. But even if x86 processors are used, the cost of R&D for an SMP memory backplane is high, and those backplanes lag behind standard high-performance interconnects. There's also a greater cost involved. Those products are proprietary, and the end users are usually locked into a specific computer manufacturer's hardware architecture, sometimes even a proprietary software stack.

X86 SYSTEMS

x86 architecture was originally designed for personal computers, and has evolved to provide low-cost server solutions with up to eight processors in a "glue-less" setup (for Intel processors) or dual-processor (for AMD). It delivers the best price-performance ratio for server systems within that class. However, x86 architecture poses unique challenges for building larger SMP systems, since it lacks some important attributes for scaling the system and increasing its memory footprint.

In addition, the technology refresh cycle (processors, peripheral devices and components) is 12 to 18 months, while new high-end SMP systems typically take three years to design and build. This mismatch creates significant risk for computer manufacturers in designing systems. It's difficult to amortize and recoup the R&D investment for rapid market changes. That's why there are very few x86-based, scalable SMP system models on the market.

Further, in the Intel case, Intel mandates that systems of either processors or more use only high-end processor part numbers, which cost significantly more per compute unit. This further reduces the allure of Intel-based SMPs to end users.

SYSTEM UTILIZATION

In many cases, SMP systems are used only for part of an HPC workflow. As some SMP codes do not scale well, it's hard to keep the system fully utilized all the time. Further, some applications only need the large memory of the SMP, but not the CPUs, yet to add more memory on those SMPs one must add processors. Therefore, customers who invest in these expensive SMP systems many times find that their SMP systems — the most expensive resource in their data center — is underutilized.

What's even more frustrating is the effect of an SMP forklift upgrade on the end user. When the user grows out of peak capacity, it makes no sense to expand the system with CPUs from the older generation, and the new CPUs do not fit the aging machine. Thus, the user who has purchased these systems for peak usage based on an estimated load three years ahead finds themselves retiring an old system only to spend a lot on a new one.

WHAT'S NEEDED IN AN SMP SOLUTION

Any solution to address traditional multi-processor shortcomings must meet the following requirements:

RUNS APPLICATIONS DESIGNED TO LEVERAGE SMPS

The customer ought to be able to run different types of applications without requiring advanced, resource-management tools. Such applications include:

- Multi-threading
- Multi-process throughput (no messaging between processes)
- Distributed multi-process (such as MPI applications)
- Large-memory applications (single-or multi-threaded)
- Data-intensive applications

The optimal solution should provide customers with the flexibility to run these different applications without complex reconfiguration or system setup. For example, a good solution would enable using the same compute infrastructure to concurrently run multi-threaded applications with throughput jobs, or be used for both distributed applications (needing high-memory bandwidth) as well as large-memory applications (needing many terabytes in memory footprint).

ENSURES PERFORMANCE THAT EQUALS OR SURPASSES SMPS

The appropriate solution should scale its performance to match traditional SMP solutions across compute, memory and I/O resources. Additionally, it should provide tools that help software engineers optimize application-level performance and scalability, leveraging the solution's architecture and system resources.

LEVERAGES THE LATEST GENERATION OF PROCESSORS AND INTERCONNECTS AT ANY POINT IN TIME

The solution should be designed to tap into the x86's fast technology refresh cycle. It must allow customers to rapidly incorporate commodity components, shortening the new product design cycle. It must also allow computer manufacturers to plan and recoup their investments quickly.

ALIGNS MANAGEMENT COSTS ACCORDING TO SMP DEPLOYMENT MODEL

The solution should match SMP systems' simple operational model, a single point of management for easy implementation and ongoing operation. This approach reduces system management overhead and contributes to lower TCO, and simplifies the deployment and management of small to mid-size HPC clusters.

MINIMIZES ACQUISITION COSTS AND CUSTOM HARDWARE USAGE

The solution should be based on industry-standard components (computers, interconnects) to leverage economies of scale, reducing overall cost.

AGGREGATION: A NEW VIRTUALIZATION PARADIGM

WHAT IS VIRTUALIZATION?

Computing virtualization is a technique for hiding a compute resource's physical characteristics from the operating system (OS), applications or end users interacting with that resource. Different types of computing virtualization paradigms are in use by IT. Examples include **server virtualization**, a single physical server appears to function as multiple logical (virtual) servers **(virtual machines)**. This kind of virtualization can also be defined as **partitioning**. Another example is **desktop virtualization**, the physical location of a PC desktop is separated from the user accessing the PC. This remotely accessed PC can be located anywhere, typically a virtual machine **(VM)** in a data center, while the user is located elsewhere. This kind of virtualization is also known as **remoting**. The last example is **infrastructure virtualization**, a set of physical servers implementing a software-defined IT infrastructure that virtualizes all of the elements of conventional "hardware-defined" systems, which includes, at a minimum, virtualized computing, a virtualized SAN and virtualized networking. This kind of virtualization can also be defined as **hyperconverged infrastructure (HCI).**

ScaleMP has created a new type of computing virtualization paradigm:

High-end virtualization: Multiple physical systems appear to function as a single logical system. ScaleMP defines this virtualization paradigm as aggregation, the inverse of partitioning.

The groundbreaking vSMP architecture aggregates multiple x86 systems into a single virtual x86 system, delivering an industry-standard, high-end symmetric multiprocessor (SMP) computer. In 2003, ScaleMP introduced patented technologies that use software to replace custom hardware and components, presenting a novel, revolutionary computing paradigm.

THE VERSATILE SMP (VSMP) ARCHITECTURE

Patented, time-proven Versatile SMP (vSMP) architecture enables users to create high-end SMP systems. The vSMP architecture replaces the functionality of custom and proprietary chipsets with software and commodity interconnects such as InfiniBand. It utilizes only a tiny fraction of the system's CPUs and RAM to provide chipset-level services, without sacrificing system performance.

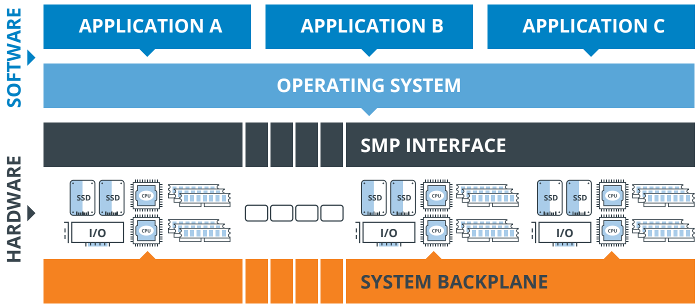
> "By providing a single virtual system, the IT complexity is significantly reduced, and while having access to large shared memory for our most demanding larger simulations when needed."
>
> Dr Carol Roberts, Research Fellow, Automotive Engineering Applied Research Group, Coventry University

vSMP Foundation is ScaleMP's implementation of the vSMP architecture. vSMP Foundation aggregates multiple x86 computer systems into one larger SMP system, allowing system vendors and value-added resellers to create high-end x86 solutions using industry-standard components, eliminating the need for lengthy and onerous custom hardware development.

Below is an explanation of a traditional SMP system, followed by a description of vSMP architecture.

LEGACY (MONOLITHIC) SMP SYSTEM ARCHITECTURE

SMP systems run a single OS. The OS interacts with the system hardware using a well-defined hardware interface, providing the OS with predefined services to use and control the hardware. These interfaces may include hardware detection and probing, memory ordering semantics, I/O space access and interrupt delivery mechanisms. An example of such hardware interface would be Intel's MultiProcessor Specification, as follows:



System boards	Processors	Operating System	Backplane	SMP Interface
Always Proprietary	Commodity: Xeon, EPYC Proprietary: SPARC, POWER	Commodity: Linux (Xeon, EPYC, Power) Proprietary (AIX for Power, Solaris for SPARC)	Always Proprietary	Commodity: Intel MP Spec Proprietary: Per each IHV

The MultiProcessor Specification (MP Spec) ... defines an enhancement to the [x86] standard to which system manufacturers design DOS-compatible systems. ... the MP defines a standard way for the operating system to communicate with the hardware. The existence of a standard interface between the hardware and the OS makes it easy for the OSVs and OEMs to quickly support a wide range of platforms with one OS version, a benefit they already enjoy in the uniprocessor desktop market for Intel Architecture CPUs. In essence, the MP Spec brings the same "shrinkwrap" benefits of the desktop market to the MP market.

MP-capable operating systems will be able to run without special customization on multiprocessor systems that comply with this specification. End users who purchase a compliant multiprocessor system will be able to run their choice of operating systems.

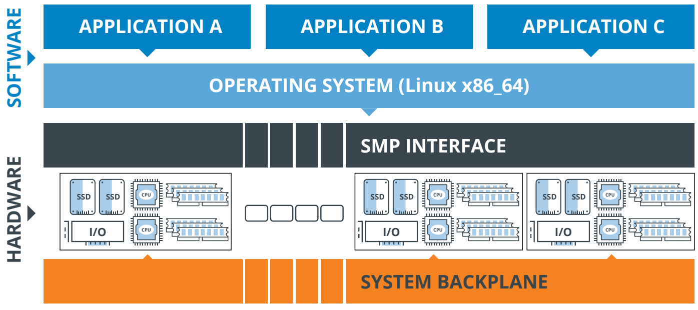
Intel's MultiProcessor Specification (and its enhancements) allows a single copy of an OS to run on a single CPU system as well as on a multi-CPU system with thousands of CPUs. It details a well-defined interface that allows the OS to probe the hardware, identify the underlying system, and then operate as it should. This interface also coordinates the underlying system with the OS. In an SMP system, the interface is implemented by a firmware and a silicon chipset.

The proprietary legacy systems' closed architecture and high R&D costs create highly proprietary systems with varying system architectures, OSs and applications, all driving higher costs and vendor lock-in for IT organizations, as described here:

- An SMP system's CPUs, memory and I/O subsystems are all connected by a proprietary backplane or interconnect, delivering high-speed access between CPUs, memory and I/O. Examples of such backplanes are Intel's QuickPath and UltraPath Interconnect (QPI/UPI), AMD's HT (Hyper-Transport) and Infinity, Fujitsu's CrossBar for SPARC, HPE's (SGI's) NUMALink, and IBM's X-Bus for Power9.

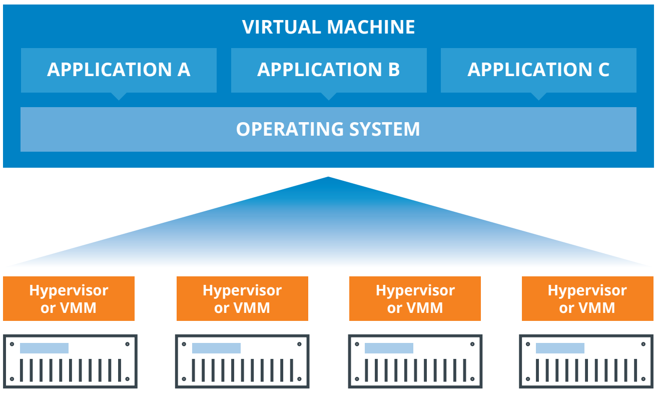
  The proprietary backplane (system interconnect) is where SMP systems differ the most from each other and where major costs of a high-end SMP system are incurred. The expenses derive from the complexity of adding more processors and connecting them to ensure both coherency and performance.

- A hardware-only SMP system requires a custom chipset to implement the system interconnect that enables processor, memory and I/O communication. The larger the system, the more complex the chipset solution.



System boards	Processors	Backplane	SMP Interface
Commodity: any vendor Dell, Lenovo, HPE, Fujitsu, Supermicro, etc.	Commodity: Xeon, EPYC	Commodity: InfiniBand, PCIe, CLX	Commodity: Intel MP Spec

In the x86 ecosystem, up to eight processors can be connected without custom chipsets, and such solutions are available off-the-shelf. x86 chipsets that support more than eight processors, however, are complicated to design. Very few implementations exist, and are limited in size to 32 processors. Moreover, since the x86 technology refresh cycle is 12 to 18 months, chipsets and boards require significant ongoing investment to keep up. This inevitably slows technology adaptations in the high-end x86 market and requires more expensive, lower-performing systems.



THE VSMP ARCHITECTURE – SOFTWARE-BASED SMP

The vSMP architecture uses off-the-shelf components. It does not require any custom parts. Its key value is the use of software to provide the functionality that is otherwise provided by a chipset, found in traditional multi-processor systems.

vSMP Foundation provides cache coherency, shared I/O and the system interfaces (BIOS, ACPI) that the OS needs. The vSMP architecture is implemented completely transparently. Users need no device drivers, no OS nor application modifications.

VSMP FOUNDATION AGGREGATION PLATFORM

vSMP Foundation requires:

- Multiple, industry-standard x86 systems with processors and memory (varying processors and amounts of memory across boards are allowed).
- InfiniBand fabric interconnect in the form of HCAs, cables and switch (switch not mandatory for some configurations).
- vSMP Foundation software, loaded from a flash-memory device (such as physical or virtual USB memory drive) or from the network.

Each system must load vSMP Foundation upon boot, by configuring its boot device appropriately.

One System

Once loaded into the memory of each of the computers making up the virtual SMP (system boards), vSMP Foundation aggregates the compute, memory and I/O capabilities of each computer and presents a unified virtual system to both the OS and the applications running on top of the OS. vSMP Foundation uses a software-interception engine — a virtual machine monitor (VMM) — to provide a uniform execution environment. vSMP Foundation also creates the required BIOS and ACPI environment to provide the OS (and the software stack above the OS) with a coherent image of a single system.

Coherent Memory

vSMP Foundation maintains cache coherency between the individual system boards and manages the memory of the virtual SMP by using multiple advanced memory management and coherency algorithms. These complex algorithms from the domains of real-time machine learning and of cache management operate concurrently on varying block sizes based on real-time memory activity access patterns. vSMP Foundation leverages board local memory together with caching algorithms to minimize the effect of interconnect latencies, and implements a software memory management unit (SW-MMU), which has mixed COMA and NUMA attributes.

Shared I/O

vSMP Foundation aggregates I/O resources across all boards into a unified PCI hierarchy and presents them as a common pool of I/O resources to the OS and the application. The OS can utilize all the system storage and networking controllers toward providing high-I/O system capabilities.

Versatile System

vSMP Foundation aggregates heterogenous system boards with different processor models, varied memory amounts or dissimilar I/O devices.

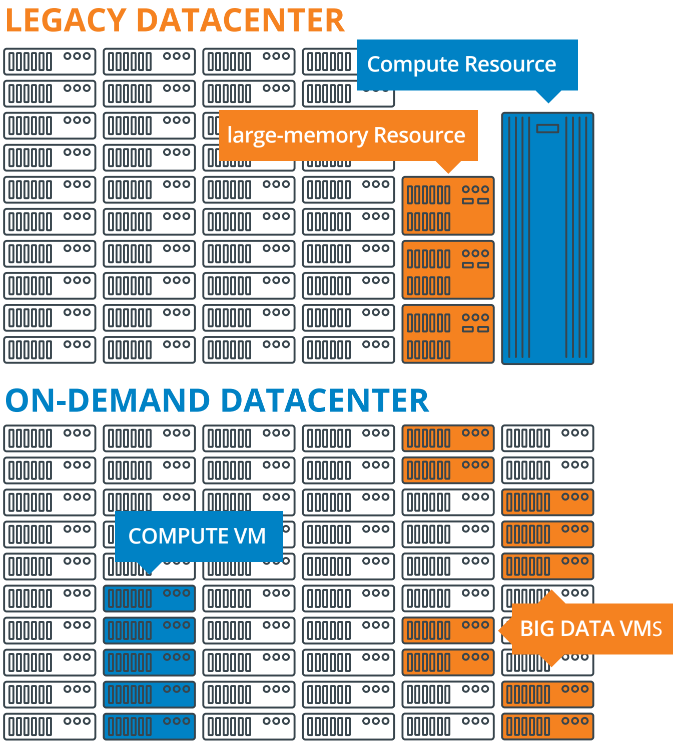
This capability is unique among x86 shared-memory systems.

For compute-intensive applications, users can architect a homogeneous system with up to 1,024 processors (more than 30,000 cores) and 2PB RAM, delivering more than 1 PFLOPs.

Selective Scaling

For memory-intensive applications, users can architect an imbalanced configuration using both high-speed and low-speed, lower-cost processors. If so configured, vSMP Foundation will expose to the OS only the top-performing processors while hiding the rest of the processors, essentially turning the off to only be used as memory controllers. This configuration reduces acquisition costs by allowing the users to buy only what they need, while also reducing operating expenses through lower cooling and power consumption, providing large memory and top system performance.

THE VIRTUALIZED DATACENTER



Similarly, the customer can mix and match I/O expansion options to fit application needs, enabling delivery of the industry's most versatile and flexible high-end x86 systems. Coupled with price/performance attributes, vSMP Foundation-based solutions provide customers the best value for their money.

On-Demand SMP

Many HPC data centers provision different computing resources for different classes of workloads; for example, a cluster with a high-end fabric for parallel-distributed jobs (MPI) and for throughput jobs (HTC), fat nodes for large-memory jobs, and SMPs for scalable multi-threaded jobs (e.g. using OpenMP). This leads to complexity and underutilization, as specific resources are only busy when relevant jobs enter the queue, and are otherwise idle. With SMPs, the situation worsens, as the SMP resource is acquired for a predicted maximal scale, but typically is shared between smaller mid-size jobs, adding to management complexity.

With vSMP Foundation, customers no longer need to buy different resources. All computers in the data center can be similar, using the most effective FLOPs/$ systems, which are typically the dual-socket workhorse. Larger resources can be created on demand (e.g. if a scalable OpenMP job enters the queue, 16 nodes can be aggregated on demand to form a shared-memory system with more than 1000 CPUs; if a large-memory job enters the queue, 10 nodes can be aggregated on demand using selective scaling to have the memory of all 10 nodes but the CPUs of only one node, etc.) The advantages of this on-demand capability are many:

- Lower overall cost, or higher FLOPs per given budget
- Increased utilization through the assignment of same hardware class to many or all application classes
- Easier management and maintenance through standardization of hardware models in the data center
- No "forklift upgrade" — the SMP is a logical function that can be used on future cluster hardware immediately after hardware refresh
- Isolation of SMP jobs — no longer needing to share machines as each job gets "its own SMP," no inter-influence between jobs

VSMP FOUNDATION AGGREGATION PLATFORM

SCALEMP IS A PROVEN, LOW-COST SOLUTIONS TO SCALE-UP X86 SYSTEMS

VSMP ServerONE

vSMP ServerONE aggregates multiple, industry-standard, x86 servers into one single virtual high-end system, serving as a superior alternative to expensive legacy multiprocessing (SMP or NUMA). With vSMP ServerONE, computing and memory are no longer linearly tied, and users can selectively scale different system attributes.

vSMP ServerONE turns multiple servers into a single computer system seen by the operating system, applications, administrators, developers or users as a single entity, running only one copy of the operating system. Each CPU in the aggregated system has access to all the memory, enabling applications to scale using thread-parallel execution models such as OpenMP. vSMP ServerONE functionality can be delivered on-demand, turning any collection of nodes connected to the same fabric into a single system, and providing a true software-defined composable server infrastructure for HPC and enterprise data centers.

VSMP ClusterONE

vSMP ClusterONE turns an HPC cluster into a unified system running a single operating system. Instead of trying to ease cluster management, it eliminates the cluster altogether. In other words, if you are worried about making the leap from a desktop or workstation to server-based HPC, and wish you could just buy a "workstation on steroids," vSMP ClusterONE gives you exactly that.

vSMP ClusterONE creates a super-sized workstation and enables the leap from desktop computing to HPC-scale deployment without any special knowledge requirements. This effectively eliminates any human-resource overhead for cluster management, and enables any Linux system admin to manage the system effectively without requiring special expertise in system provisioning, cluster management, fabric management or parallel file systems. All capabilities are abstracted and available as standard Linux OS functions on the aggregated system.

WANT MORE INFO? NEED TO SEE IT TO BELIEVE IT?

Need additional technical information, system requirements or want to schedule a live, hands-on demo of vSMP Foundation?

Visit our site **www.scalemp.com**
or mail **info@scalemp.com**

ScaleMP™
Creating the Power of One