uniserver

# A Universal Micro-Server Ecosystem Exceeding the Energy and Performance Scaling Boundaries

www.uniserver2020.eu

**UniServer facilitates the advent of IoT solutions through the adoption of a distributed infrastructure where decisions are taken more locally at the Edge of the Cloud with novel energy-efficient micro-servers.**

## On the Verge of the Free Technology-Scaling Ride End

In the past few decades, aggressive technology scaling was the main driving force for the explosive performance improvement of digital systems that radically reshaped the ways that we work, entertain and communicate. New paradigms such as cloud computing and edge computing and the Internet of Things (IoT) enable intelligent inter-connection of billions of devices. These devices will generate huge volumes of data (exabytes) that need to be processed and analyzed in centralized or de-centralized data-centers close to the users. The analysis of such data is expected to lead to new scientific discoveries and new applications that will further improve our everyday life.

However, such paradigms are at threat since the “free-ride” of technology-scaling seems to come to an end with nano-scale circuits becoming very prone to potential failures due to increased static and dynamic variations of circuit parameters. In particular, as devices are being pushed to the atomic scale, it is becoming very difficult to fabricate circuits with the expected performance and power consumption. Consequently, many fabricated chips will not meet the specifications since they are ‘intrinsically different’ thus putting at risk the correct functionality of products. Under this reality, manufacturers insist on obeying traditional deterministic error-free computing rules by hiding the huge performance/power variability in the fabricated chips from the software layers by adopting pessimistic ‘safety’ timing margins and redundant (thus performance- and power-hungry) error-correction schemes. However, such safety margins end up prohibiting circuits and systems to work as fast and more importantly as energy-efficiently as they could.

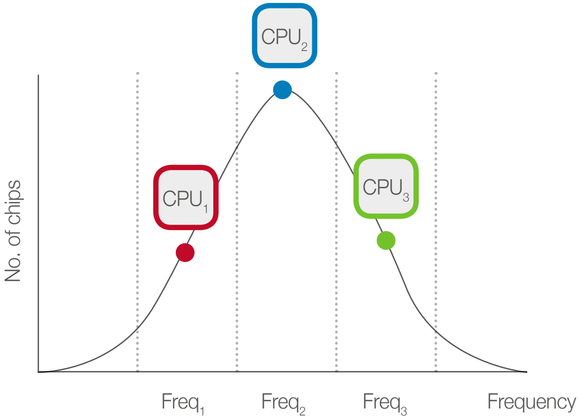
# Substantial Improvements Require Radical Rethinking of System Design

In order to substantially improve the energy-efficiency there is a need to design new error-resilient server ecosystems that are able to deal with the increased hardware variability in a more intelligent way than the conventional pessimistic paradigms.

UniServer project turns the table around and puts forth the following question:

> Why allow the worst operating margins of fabricated chips to artificially constrain the performance and energy of today systems?

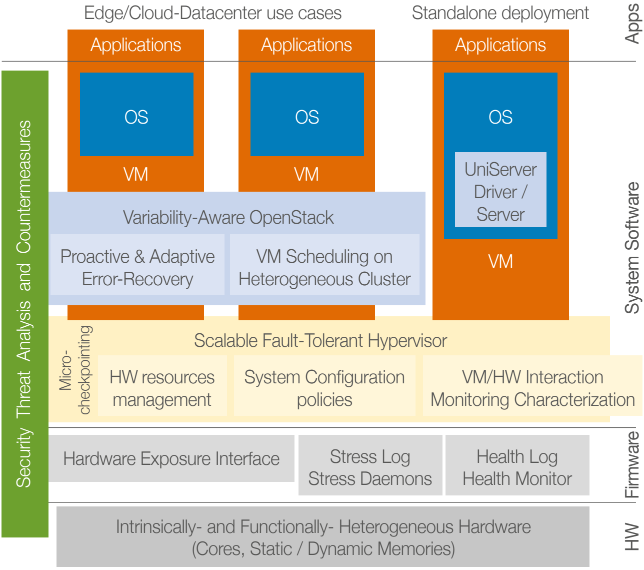
The reality is that each manufactured processor and each memory module is inherently different and lies on a distinct performance. Putting it simpler each of the manufactured chips may be able to operate at different frequencies (preferably higher) and supply voltages (preferably lower) that may lead to more energy efficient operation.




In the UniServer, we believe that the computing industry needs to see such heterogeneity not as a problem but as an opportunity to improve energy-efficiency especially in next generation servers. ‘Functional heterogeneity’ has already been adopted in embedded systems and servers with hybrid CPU/GPU/accelerators architectures. Therefore, it is now time to also expose the ‘intrinsic heterogeneity’, harness it and use it to our advantage by redesigning the hardware and software for improving energy-efficiency or performance.

> Aim at revealing new energy efficient voltage/frequency/refresh-rate states for different voltage domains of processors and memories.

To make the most out of this opportunity, on the one hand, there is a need for developing automated procedures at the firmware level for exposing the real capabilities of each processor and memory resource. Such procedures consist of diagnostic and health monitoring daemons that will be evaluating the operation of the underlying hardware components at various operating points, beyond the nominal/conservative ones. These procedures will be collecting, mining and analyzing various parameters of interest, such as correctable and uncorrectable errors, performance counters, system crashes/hangs, thermal and power behavior. The results of such on-line system analysis and other information that will be collected on-the-fly need to be communicated through a novel Hardware Exposure Interface that communicates to the software stack the identified operating points as new attributes of the underlying hardware components. The above novel procedures will aim at revealing new energy efficient voltage/frequency/refresh-rate states for different voltage domains of processors and memories.




On the other hand, the design of a number of system software layers adopted in today's data-centers, namely hypervisors and cloud management frameworks such as OpenStack, need to be re-thought. Operating aggressively and close to the hardware limits may inadvertently introduce erroneous hardware behavior in both the storage elements and the logic of the processor and the memories. The hypervisor is the first line of defense against such faults. It should therefore be error-resilient itself, and at the same time transparently and efficiently offer the illusion of a fully reliable hardware to the edge services implemented within virtual machines on top of it. The cloud management framework has similar configuration and management opportunities, exploiting reliably the intrinsic heterogeneity, however not at the node but rather at the cloud granularity.

# Impact: Empowering Internet's Evolution

Overall, we aim at integrating the software and hardware ecosystem described above in a server prototype by 2019, which could improve the energy efficiency of running IoT and Big Data applications by 31x by 2019 based on estimations made by the UniServer consortium. This improvement will be for the same performance and relative to the energy dissipation of a commercial micro-server platform and system software circa 2014.

The realization of the envisioned energy efficient servers with a complete software stack provides an opportunity for developing a more sustainable Internet. Presently, most of the processing and storage in the Internet is performed in the cloud with massive data-centers that are located in remote locations and occupy an area of a few football fields, contain tens of thousands of servers, consume electricity of a small city and utilize expensive cooling mechanisms. Such data-centers will not be sufficient in the IoT era, but a combination of classical data-centers on centralized remote locations along with new data-centers at the edge, closer to the users, is believed to be a more viable solution. Such edge data-centers, supported by the developed micro-server could enable the widespread use of IoT devices and the development of new applications for smarter working and living environments.

European Commission
Horizon 2020
European Union Funding
for Research & Innovation
Funded under the
grant agreement
No 688540

CONTACT US
**Project Coordinator:** Dr. Georgios Karakonstantis
ECIT Institute, Queen's University Belfast, UK
**Email:** uniserver-info@qub.ac.uk

FOLLOW US
www.uniserver2020.eu/ @uniserver2020/ @uniservereu

## Members of the Project Consortium

Queen's University Belfast
www.qub.ac.uk

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
en.uoa.gr

University of Cyprus
www.ucy.ac.cy

Meritorius Audit Ltd
www.meritorius.com

UNIVERSITY OF THESSALY
www.uth.gr

applied micro
www.apm.com

arm
www.arm.com

IBM
www.ibm.com

*Sparsity
www.sparsity-technologies.com

WORLD SENSING
www.worldsensing.com