Reinventing Real-Time for Supermobile

New Challenges Demand New Solutions

Reinventing Real-time for Supermobile

Following a century of broadly linear innovation from fixed voice and data through to their mobile successors, the telecommunication industry is entering a profoundly disruptive phase often referred to as "Supermobile."

A companion whitepaper produced by MATRIXX Software, *Competing in the Supermobile Era*, examines Supermobile from a market viewpoint, exploring associated changes in market landscape, opportunities, and customer expectations. The paper concludes that communication service providers (CSPs) are at a cross-roads and must re-invent themselves *now* to exploit the exciting opportunities offered by Supermobile. To delay is to invite substantial, long-term, and potentially irreversible decline.

This whitepaper focuses on the growing importance of **real-time technology** in enabling the service capability, agility, and transparency essential for a differentiated Supermobile experience. We argue that existing approaches, particularly as applied to online charging and policy, are fundamentally limited in their ability to meet Supermobile performance and capability demands. CSPs are already forced to make deep trade-offs between capability, performance, predictability, and affordability with serious consequences for innovation, growth and ultimately survival.

These limitations arise principally from the use of real-time techniques little changed in decades. Data integrity is assured through Online Transaction Processing (OLTP) principles introduced by IBM nearly 40 years ago. Fundamental to OLTP is the use of database locking to ensure transactions are successfully committed or cleanly aborted. As we will discuss, locking can introduce huge performance penalties for Supermobile transactions, characterized by extreme throughput, very low latency targets, and increasing complexity.

Online charging solutions are frequently based on evolution of Intelligent Network (IN) technology originally developed around the relatively simple and predictable demands of circuit-switched voice. These solutions are struggling to offer the service agility, flexibility, and velocity demanded by data-centric Supermobile applications and services.

Supermobile creates fundamentally new challenges for real-time technology. It requires fundamentally new solutions, not optimization and architectural tinkering. Relying on old thinking and old technology to solve the new challenges of Supermobile amounts to what Einstein might have called *insanity*: "Doing the same thing over and over again and expecting different results."

MATRIXX Software has responded to these challenges with the development of **Parallel-MATRIXX™** technology, a fundamental reinvention of real-time transaction processing targeted at the extreme demands of Supermobile. Parallel-MATRIXX™ technology creates a new real-time sweet spot allowing CSPs to pursue exciting Supermobile business opportunities, free from the painful trade-offs described earlier.

This paper examines why Supermobile needs real-time, how the particular characteristics of Supermobile are stretching contemporary real-time technologies beyond breaking point, and how Parallel-MATRIXX™ technology overcomes these limitations. Benchmarking results from a TM Forum Catalyst Project sponsored by Orange and Swisscom are presented, confirming performance and efficiency improvements greater than 100x relative to legacy approaches, with a "fraction of the hardware" and with no performance degradation for even the most complex Supermobile scenarios.

Supermobile and the Need for Real-Time

Supermobile, an era when mobile communication technology transcends simple person-to-person voice communication, instead offering a richly featured real-time interface between individuals and their digital world and where voice is no longer 'the service' but simply a feature. An enabler of multimedia communication, collaboration, and entertainment. A medium for stimulating, enabling, and globally diffusing innovation, new business models and economic value at the speed of light. Disruption at its most profound. Exciting times – and we're only on the starting line.

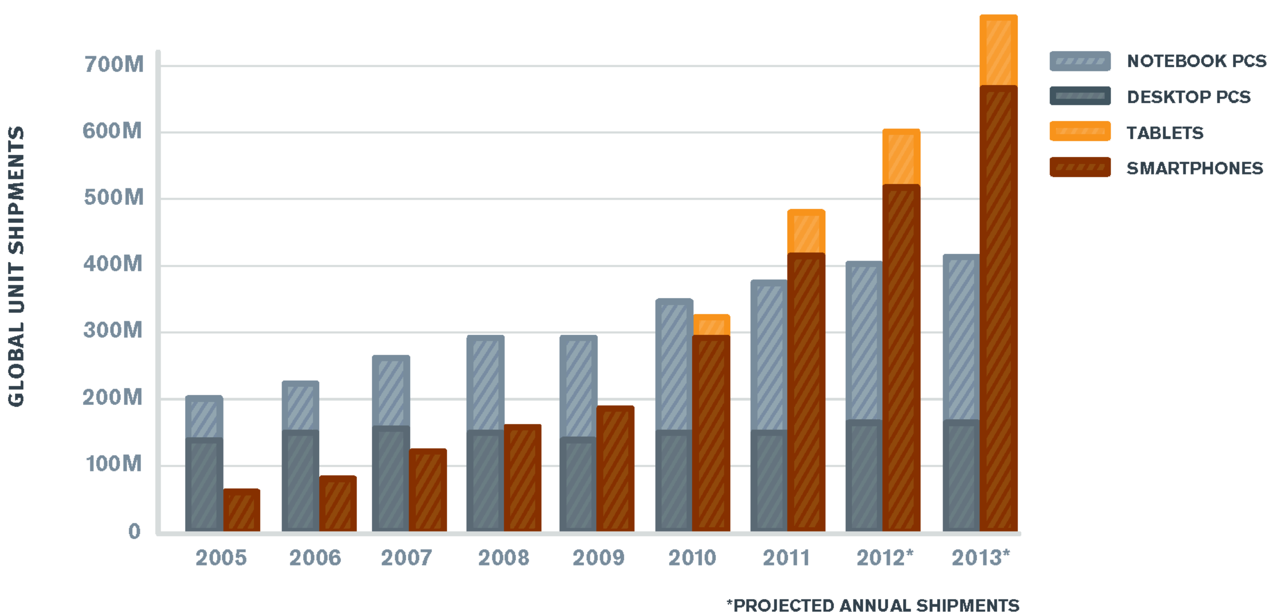
As with many disruptions, the tipping point is triggered through a convergence of trends. For Supermobile, we identify four such trends that combine to yield fundamentally new challenges for real-time technology, namely Devices, Long Term Evolution (LTE), Creating the "Online" Experience, and Service Ecology.

Devices

Android™ OS has a codebase estimated at 10 million lines[i] and was installed on approximately 70% of Smartphones sold during the second quarter of 2012. By 2013, it's estimated that smartphone sales will exceed non-smartphone sales,[ii] and by 2016 the typical customer in mature markets will use three communication-enabled devices.[iii]

Such devices can no longer be considered as mere phones. They are high-performance computing platforms integrating a rich array of sensors, network interfaces, APIs, and high-resolution, touch-sensitive displays. Simply put, they are becoming the means by which humans interface to and engage with their digital environment for work and pleasure. The growing sophistication and diversity of applications and services on these devices reflects this changing role with significant consequences for CSP network and IT systems.

An early example of these consequences arises from smartphone applications creating an increasing number of short-burst background sessions for synchronization (email, cloud storage, etc.). In the not uncommon case that the recipient server for these is overloaded, offline for maintenance, or simply gone, the smartphone application is nothing if not tenacious. It will continually retry along with thousands or even millions of other clients globally that are also trying to reach the same server. These behaviors have profound implications for real-time systems attempting to impose charging and policy on these sessions.



By 2013, sales of smartphones and tablets together are expected to more than double the total sales of desktops and laptops. (Source: "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010 – 2015.")

Long Term Evolution (LTE)

Accelerating LTE deployment is witnessing an increasing number of CSPs offering throughputs approaching 100Mbps.[iv] LTE-Advanced will offer throughput theoretically exceeding 1Gbps for slow moving users – realistically, most smartphone users. While the need for speed makes great headlines, its ugly sister, "packet round-trip time or latency," can be more important for delay-sensitive applications such as Voice over IP (VoIP), gaming, and generally any TCP-based applications requiring multiple round trips to ramp-up throughput.

Early mobile packet technologies such as GPRS offered poor latency, commonly exceeding 100ms, preventing their use for a large class of delay-sensitive applications. With LTE driving latencies towards 10ms, these applications become feasible. They remain, however, highly sensitive to latency degradation due to traffic congestion or signaling bottlenecks in, for example, real-time charging or policy control servers.

Creating the 'Online' Experience

Consumer expectations are increasingly influenced by their online experiences characterized particularly by immediacy, personal relevance, and community.

Amazon™ offers an excellent example of these principles in action. Personalized product offerings reflect purchase and browsing history. A recommendation engine drives confidence and loyalty through transparency and collaboration with a community of shared interests, and the 1-Click® purchase option allows simple and immediate ordering, payment, and fulfillment. For electronic books, delivery to the Kindle™ is complete in less than a minute of the purchase. Spend history and order status is immediately available and accurate.

The same principles should be coded into the DNA of Supermobile offerings. The effectiveness and efficiency of Supermobile CSPs in building and leveraging rich user profiles combining *core profile* (demographics, calling history,…), *preferences* mined particularly from mobile payments and partners, and *context* (location, retail partner proximity, connected device type, and so on) will be crucial to differentiation and growth. For example :

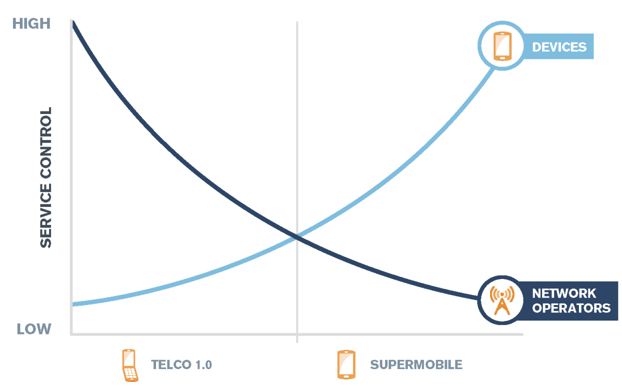
- A retail subscriber whose profile indicates regular use of coffee shops can be presented with an ad-hoc promotion when in the proximity of a partner outlet. Mobile payment to purchase a beverage credits a loyalty scheme and immediately forwards a link to the subscriber allowing 30 minutes of high-definition access to a set of carefully selected movie promotions.

- A business traveler can be offered discounts for partner restaurants based on food tastes (either explicitly configured or "mined" from previous purchases), one-click reservations, and route-planning or taxi booking.

While there are doubtless many more creative and valuable use cases, these simple scenarios underline the need for highly integrated real-time charging, policy, and analytics/personalization capabilities.

Service Ecology

From circuit-switched voice (predictable in growth and behavior), through SMS (predictable in behavior but perhaps less so in growth), to a continuously and rapidly evolving ecology of millions of communicating over-the-top (OTT) applications and services predictable only in their unpredictability. This is the Internet model: service execution and intelligence in the end-points, interconnected via a resilient, adequately dimensioned network.



Supermobile taking control

The decoupling of application or service development and execution from the network creates fundamentally new challenges. No longer can CSPs rely on standards adherence to assure end-to-end integrity and performance through careful specification of application and network interaction. Earlier we presented the example where OTT applications can create a storm of re-try requests in the event of a server outage. This is typical of the challenge arising from this new distributed service model, and one with potentially serious consequences for any real-time charging and policy servers that happen to get in the way.

The service challenge for Supermobile CSPs is achieving OTT levels of innovation and agility while assuring traditionally demanding network availability targets. Whether such capabilities are deployed as a means of OTT collaboration, competition, or some completely new direction is a strategic decision for individual CSPs. Recent examples of collaboration include Spotify access in some T-Mobile™ plans[v] and Whatsapp Messaging offered by Smartone.[vi]

The Case for Real-Time

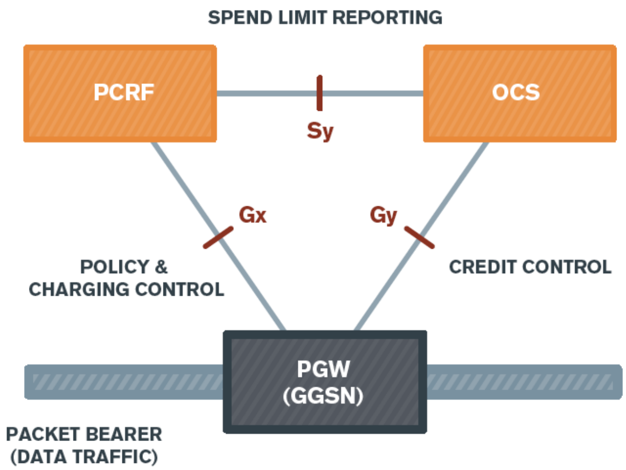
Considering the Supermobile characteristics and challenges outlined above, we can identify a spectrum of capabilities dependent on, or heavily benefiting from real-time:

1. Prepaid charging – where it all began!
2. Cost control supporting CSP credit risk management and regulatory bill shock compliance.
3. Policy control ensuring correct allocation of network resource (bandwidth, delay, priority, etc.) for multimedia services.
4. Fair Usage Policy (FUP) imposing penalties such as overage charges or QoS downgrade for data consumption beyond a periodic usage cap.
5. Supporting payment, access type, and service convergence through deployment of a common rating engine. A barrier to convergence has been the difficulty in developing a real-time Online Charging Server (OCS) that supports the rich product catalog possible with offline batch processing systems. These challenges are particularly evident with enterprise customers where charging must support large and complex hierarchies. These barriers are further symptoms of fundamental limitations in existing real-time technologies discussed in more detail below.
6. Online self-care offering customers immediate and accurate charge and usage information, service configuration, and ordering.
7. Flexibility for customers to pick and mix their own product offers, freely selecting and combining voice, messaging, data, and other service components.
8. Creation and promotion of ad-hoc offers, for example the 30-minute high-quality video bundle described earlier.
9. Real-time streaming analytics that are able to integrate multiple sources to trigger time-sensitive personalized offers.
10. As Supermobile drives enhancement of network and IT technology, CSPs' ability to monitor and manage an increasingly complex service mix must also be enhanced. For example, very few CSPs currently have anything close to a real-time view of the mix and value of services transiting their data networks.

Why Existing Real-Time Technologies Will Fail Supermobile

What is it about Supermobile that so fundamentally challenges and exposes limitations in current real-time technologies? Why will incremental innovations around hardware performance, in-memory databases, and other optimizations that have served well for many years no longer be sufficient?

To offer a tangible response to these questions, we move from real-time in a broad technology sense to focus increasingly on specific real-time capabilities within the 3GPP Policy and Charging Control (PCC) architecture and highlighted in the diagram below – namely, the Online Charging Server (OCS), Policy and Charging Rules Function (PCRF) and Packet Gateway (PGW).



3GPP PCC architecture

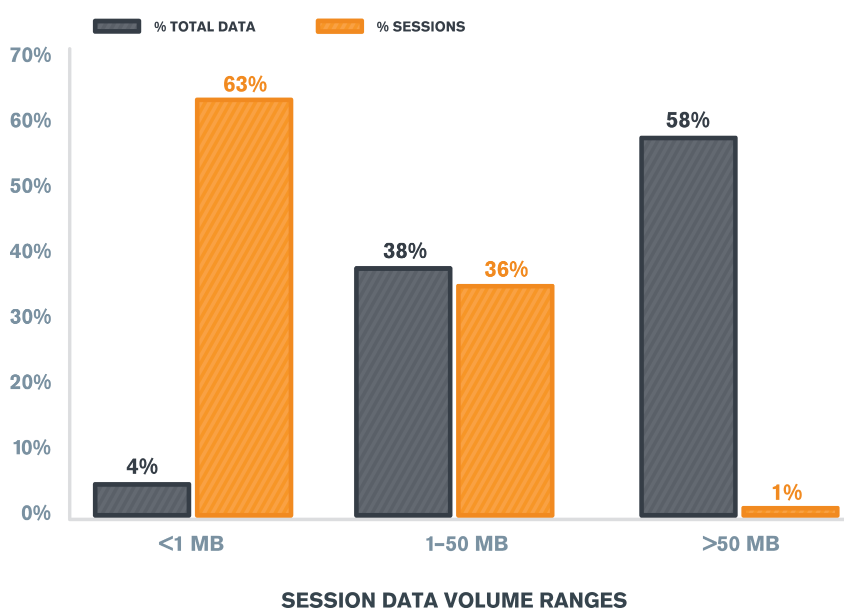
Challenge #1: Session Characteristics

Much attention is given to mobile data growth forecasts in terms of headline metrics such as total volume of traffic, smartphone penetration, etc. However, from the perspective of real-time system impact this is the tip of the Supermobile iceberg. The hidden danger relates not so much to the total volume of data, but to the characteristics and statistics of that data.

As outlined earlier, smartphone applications are driving disproportionate growth of relatively small data sessions, many associated with application background activity. The chart below is based on real CSP data and presents session statistics measured over a 24-hour period.

As the chart indicates, 63% of sessions are 1 MB or less, significant in terms of driving OCS charging transaction load but also constraining CSP options to reduce the load significantly. To understand why, a short primer on OCS quota management is useful.

Consider a simple WEB browsing session charged at 10 cents/MB.[vii] Ignoring policy (PCRF) interactions, the PGW will generally use the Diameter Credit Control Application (DCCA) protocol to request session authorization. Successful authorization returns a quota grant based on some combination of volume, time, events, or money. In this case, assume a 20 MB volume grant is returned. Continued usage consumes the grant and, shortly before exhaustion, the PGW will request further quota from the OCS, repeating this process until either the user terminates the session or a limit is reached and the OCS refuses further requests.



Frequency and volume distribution of mobile data sessions

The decision on the size of quota to be granted by the OCS is becoming increasingly complex due to the parameters and trade-offs influencing the decision. For example:

- **Throughput –** The rate at which volume quota will be consumed is broadly proportional to throughput. A simplistic analysis predicts a 2,000x increase in OCS transaction load due to quota re-authorizations if comparing GPRS (approx. 50kbps) with LTE (100 Mbps) for a fixed quota size. The implications of moving to LTE-A at 1Gbps are clear.

- **Limits –** For prepaid users or postpaid with cost control, there will be a limit where further requests for quota will be rejected. To ensure this limit is not exceeded, some solutions scale down quota grants as limits are approached.

- **Fragmentation –** Imagine our web surfer clicks on a link to set up a VoIP call or access streaming video. This could be charged at a different rate resulting in a parallel request for quota subject to different pricing (and potentially policy) rules. Such parallel flows can lead to fragmentation of quota across multiple flows, often locking quota in as those flows become idle. A related issue, charging parallel sessions, is discussed in the next section.

- **Hierarchy –** Shared balances across devices and family or enterprise groups are becoming increasingly popular and extend the fragmentation challenge due to the additional sharing dimension. Later we consider the enterprise case where a single allowance may need to be distributed across thousands of users.

Some quota management strategies assume frequency of quota requests, and hence real-time OCS load, can be reduced by configuring larger quota grants. But such an approach is only useful where the size of the grant is significantly less than the session size. In the chart presented above with 63% of sessions being less than 1 MB, increasing quota grants above this size yields no benefit. The PGW will simply retain unused quota until a timeout, forced re-authorization, or the session terminates, creating fragmentation as described above.

It could be argued that increasing quota size is more effective when applied to the small volume (1%) of sessions greater than 50 MB, or larger sessions generally. A challenge with large sessions, however, is that they are often associated with peer-to-peer file transfer and streaming video applications with huge bandwidth appetite. So even larger quota grants will be consumed rapidly at LTE throughputs.

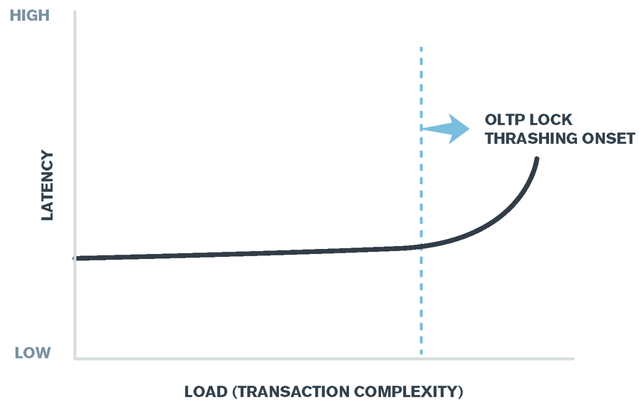
In short, sophisticated quota management algorithms are limited in their efficacy to reduce transaction load. Furthermore, the effectiveness of any strategy will be closely coupled to the mix of applications, network capacity and subscriber behaviors driving load at any given time. As this mix evolves at Supermobile speeds, so must these algorithms. It's a fast moving target and a typical example of increasingly complex optimization techniques offering diminishing returns.

It's time to face the elephant in the room. Real-time Supermobile demands will not be served by layer upon layer of optimization and endless capacity upgrade. The smart money is on understanding the fundamental limits of existing technology – and removing them.

At MATRIXX Software we believe the answer, alluded to in the introduction, lies principally with two technologies at the heart of the real-time platforms supporting transactional charging and policy applications:

- Online Transaction Processing (OLTP) techniques assuring transaction and data integrity are inefficient in the face of Supermobile throughout and traffic statistics. As transaction loads increase, latency associated with OLTP database locking rapidly becomes significant when charging and policy servers need to respond to network requests in a few 10s of milliseconds. Beyond a certain load, performance can degrade extremely rapidly due to 'lock thrashing' as illustrated below. Latency is further increased by the clustering needed to atomically replicate transactions across redundant severs in order to assure CSP availability targets.

- Additional latency is introduced by charging and policy applications which typically require execution of complex conditional logic for every transaction. As Supermobile demands more innovative and personalized offers, this logic becomes more complex, driving latency higher still.



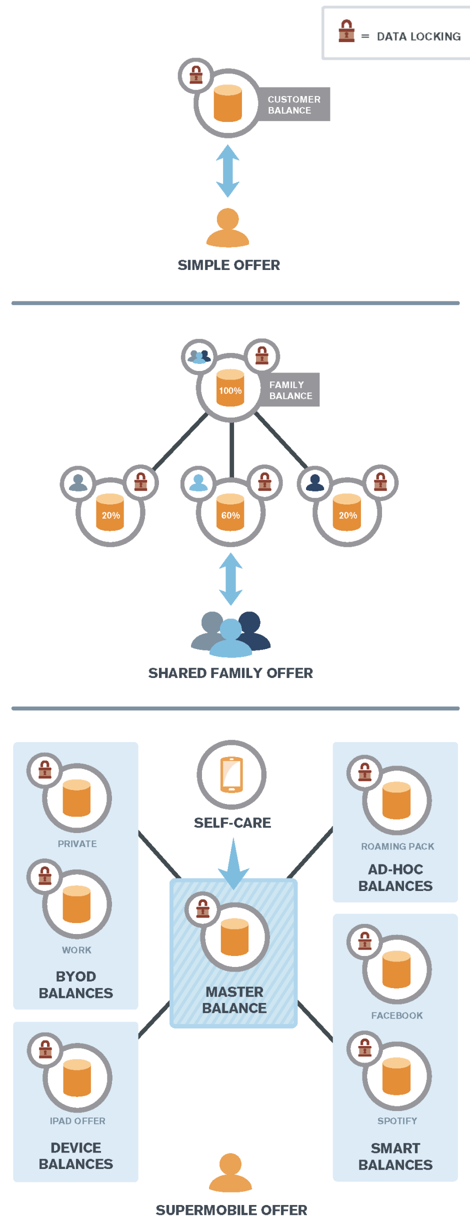
Impact of locking on OLTP performance

Even for relatively simple applications such as real-time cost control and notification, increasing throughput is already challenging existing approaches and forcing many CSPs to deploy significantly more capacity than expected. In the next section we consider how the challenge is magnified through the lens of increasing transaction complexity.

Challenge #2: Transaction Complexity

Transaction complexity increases rapidly when multiple balances are impacted by a single transaction, or conversely, a single balance is impacted by multiple transactions. Such behavior is typical of sharing scenarios and more innovative offers as illustrated below.

As an extreme but high-value example, consider an enterprise scenario where a monthly package is negotiated for 1TB mobile data. We'll focus on data, but the argument extends to voice and messaging. An enterprise administrator can distribute this allocation across different departments, down to individual employees if required (perhaps for high usage international roamers). Real-time cost control is required to ensure users/departments do not consume more than their allocation, provide real-time notifications and allow allocations to be transferred between individuals or departments.



Supermobile offers drive transaction and locking complexity

To realize this capability requires balances to be correctly updated through the hierarchy in real-time for each transaction and with potentially many thousands of transactions occurring in parallel for a large enterprise. High availability demands that transactions are also committed across redundant servers. Under these conditions, OLTP lock thrashing can bring a system to its knees.

It is this challenge in particular that has hindered adoption of real-time convergent charging in enterprise segments.

Transaction complexity is also increased by the need to charge accurately for parallel data sessions associated with, for example, multiple smartphone applications. This creates particular challenges for real-time charging solutions evolved from circuit switched voice where no equivalent parallelism concept exists.

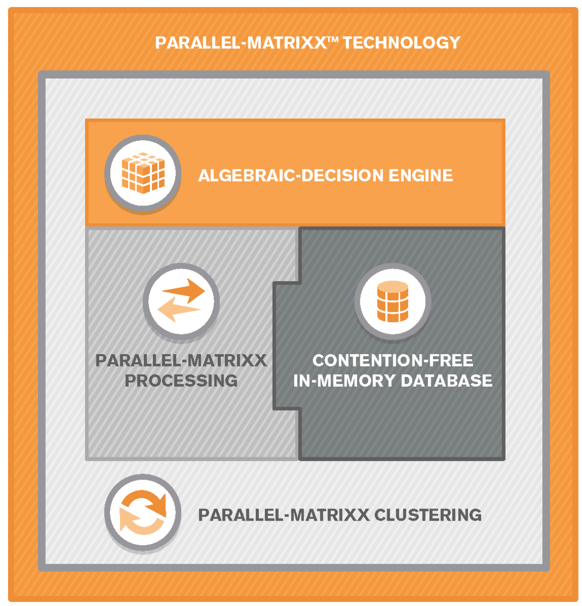
Challenge #3: Policy and Charging Integration

Fair Usage Policy was identified above in our spectrum of real-time applications. This is an important use case as it's a principle driver for the integration of real-time charging (OCS) and policy (PCRF) platforms which have traditionally followed quite separate evolution paths. OCS counters are configured to trigger a notification to the PCRF when a periodic usage threshold is reached. In response, the PCRF enforces the necessary reduction in QoS.

The OCS-PCRF integration needed to support this function has catalyzed vigorous debate regarding the evolving roles of the OCS and PCRF, particularly in respect of hosting counters, subscriber profiles, and even the rating function. This debate frequently revolves around how performance and agility can be improved by relocating functions between these components – for example, moving counters to the PCRF. As with other optimizations discussed in this paper, there is a danger of simply moving a problem rather than resolving it. We would argue that the OCS and PCRF are constrained by the same underlying real-time transactional technology limitations. Once these limitations are taken off the table through techniques described below, we believe the increasingly intimate relationship between charging and policy creates a compelling case for a unified OCS-PCRF implementation.

Parallel-MATRIXX™: Real-Time for Supermobile

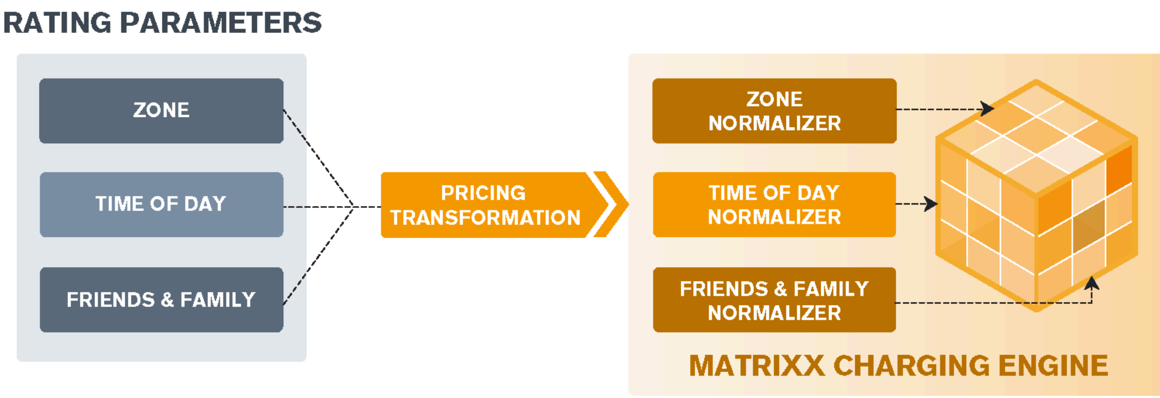
Designed to meet the demanding performance and capability requirements of Supermobile, Parallel-MATRIXX™ technology has completely re-invented transactional real-time and eliminated limitations with contemporary technologies described earlier. The diagram below identifies the Parallel-MATRIXX™ functional architecture based on multiple patented technologies, and offering a performance improvement of at least two orders of magnitude relative to legacy approaches.



The core real-time technology for MATRIXX Software

Algebraic-Decision Engine

OCS raters can be broadly classified as rule- or data-driven. The former offer great flexibility to configure rating scenarios of arbitrary sophistication but which can become challenging to maintain beyond a certain complexity. Data driven systems typically offer a rich catalog of off-the-shelf templates that are easily configured to create real offers. These templates are "baked" into code so performance can be highly optimized. The challenge with this approach arises when no suitable template is available, often requiring complex and costly customization.



MATRIXX Software normalizer concept

With respect to real-time performance, both approaches share a common weakness. Every transaction results in execution of conditional logic reflecting the rating discriminators (*if weekend*, and *if URL is On-net*, and *if…*). As rating, or indeed policy, rules become more sophisticated, execution code-paths extend and performance degrades – often unpredictably. For Supermobile with extreme throughput, high volumes of short burst sessions, and increasingly complex transactions, performance can rapidly become unacceptable.

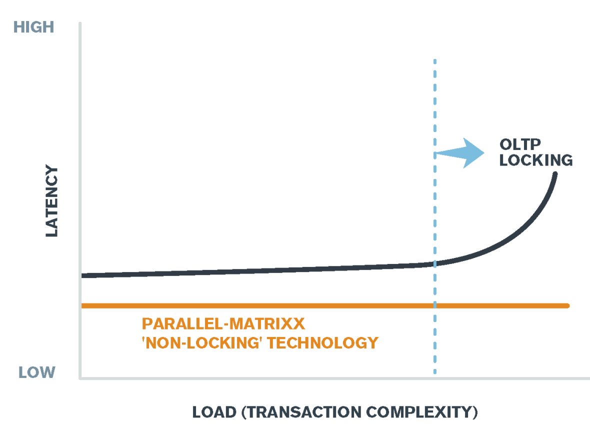
The Parallel-MATRIXX™ Algebraic-Decision engine eliminates this degradation by building on the simple principle that any pricing concept can be represented as a set of mathematical equations. Modern CPUs capable of 200 million multiplications per second are exceptionally efficient at solving such equations. Pricing plans, offers, and policies are configured via a GUI and transparently compiled into an *n*-dimensional matrix where each dimension corresponds to a rating normalizer (such as time, location, service, etc.). Stored at each matrix "intersection" is a linear equation representing the rating formula to be applied. As each transaction is mapped to the relevant intersection, solution of the associated linear equation is extremely fast.

As offers are extended with additional normalizers (for example, adding a device dependency to offer lower rates for a promoted device), the matrix dimensionality is extended accordingly. This simply results in a few additional CPU cycles to solve the rate equation with no significant impact on latency.

Contention-Free In-Memory Database and Parallel-MATRIXX™ Processing

Maintaining data and transaction integrity is a mission-critical requirement for any database containing CSP customer or financial data. For example, an attempt to transfer funds between two customers must complete successfully or be cleanly aborted. A situation where the donor's account is debited but some technical failure results in the recipient not receiving the funds would leave the database in an invalid state.

As described earlier, current real-time systems rely heavily on OLTP and locking techniques to assure data integrity but which can lead to rapidly degrading and unpredictable performance for Supermobile scenarios.



Unlike other real-time systems, Parallel-MATRIXX™ technology does not utilize locking

Parallel-MATRIXX™ technology is based on an in-memory database that does not utilize locking while still supporting full ACID-compliant[viii] transactions. No transaction is ever blocked from accessing or updating data while newly developed algorithms detect and resolve transaction conflicts. The most demanding enterprise scenarios are supported – for example, where thousands of employees may be accessing a single enterprise balance with absolutely no blocking. This innovation removes possibly the most critical performance limitation with current real-time technologies.

Parallel-MATRIXX™ Clustering

Transaction throughput and complexity creates significant performance challenges for clustering technologies that must ensure transactions are correctly committed across multiple servers. Such redundancy is essential to meet CSP demands for resilience and availability targets as high as 99.999%. Some clustering solutions in use today can only be considered partially compliant in terms of the ACID transaction model. Specifically, the need to meet very demanding latency targets results in clustering of transactions across redundant servers completing only after the network response is issued from the primary server. This falls short of the 'Atomic' ACID characteristic which demands transactions are fully committed at a *system* level before they are acknowledged as successfully complete.

To ensure the benefits gained through the algebraic rating and contention-free database innovations are not compromised by clustering bottlenecks, Parallel-MATRIXX™ also introduces new clustering technology. This ensures all transactions are clustered in real-time across redundant servers at performance levels consistent with Supermobile. By the time the network receives a response to a charging or policy request, the associated transaction is atomically replicated across all servers.

Benchmarking

A TM Forum Catalyst report can be downloaded from the MATRIXX Software website. It reports performance tests for the MATRIXX Software OCS built on Parallel-MATRIXX™ technology.

The report includes measurements of performance for a *single* COTS blade[ix] based on a scenario of 20 million customers and mixed data/voice/messaging service generating 10,000 transactions per second. With full logging, transaction response latency remained below 20ms for 99% of transactions. No degradation in performance was seen for complex rate plans, discounts, hierarchies, and balance sharing.

The report concluded that the result showed a 125x improvement in throughput and efficiency with respect to other real-time solutions and with only a "fraction of the hardware."

Conclusions

Increasing use of real-time technology is essential if CSPs are to exploit the opportunities of a Supermobile era. Contemporary real-time technologies reflect design principles established nearly 40 years ago around the simpler demands of a circuit-switched voice era. This paper describes why these principles are severely limited in their ability to address Supermobile demands. In essence, Supermobile creates fundamentally new challenges; it requires fundamentally new solutions.

Parallel-MATRIXX™ technology brings real-time up to date. Designed for a data-centric Supermobile world, it offers a new sweet spot, freeing CSPs to innovate and grow, no longer hamstrung by the need to continuously trade-off capability, performance, affordability, and predictability.

About MATRIXX Software

MATRIXX Software enables NOW. We give Communications Service Providers instant visibility, intelligence and control of services across your network. Our patented, smart charging technology enables a seamless, intuitive service experience that empowers your subscribers – in the moment, every moment. Delivering up to a 100 times increase in efficiency and scalability, we equip you with the most advanced solution designed specifically for the data generation. Charging. Policy. Insight. Value. NOW.

Follow MATRIXX Software online:

Website: www.matrixxsw.com

YouTube: www.youtube.com/user/matrixxmarketing

LinkedIn: www.linkedin.com/company/586602

Twitter: www.twitter.com/matrixx_sw

---

[i] How many lines of code in Android OS http://www.gubatron.com/blog/2010/05/23/how-many-lines-of-code-does-it-take-to-create-the-android-os/
[ii] See http://www.economist.com/node/21562939.
[iii] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010 – 2015.
[iv] Swisscom Mobile http://www.swisscom.ch/en/residential/mobile/mobile-network/4g-lte.html
[v] T-Mobile Spotify offer: http://www.t-mobile.de/spotify.
[vi] WhatsApp partnership with ‘3 Hong Kong’ http://blogs.informatandm.com/5987/whatsapps-partnership-with-3-marks-an-acceleration-in-its-engagement-with-mobile-operators/
[vii] Even for flat rate, usage (overage) charges may still apply once a periodic threshold or CAP is exceeded.
[viii] ACID: Atomic, Consistent, Isolated, Durable.
[ix] HP Proliant BL460 G6.