

# ASYNCHRONOUS AGGREGATE RESOURCE MANAGEMENT FOR MUOS

Edward K. Orcutt, Ph.D and Dean Vanden Heuvel

General Dynamics C4 Systems

Scottsdale, AZ

## ABSTRACT

*The Mobile User Objective System (MUOS) is founded upon commercial Third Generation (3G) Wideband Code Division Multiple Access (WCDMA) technology to provide voice and data services to the warfighter. An important requirement on the system is to adjudicate service access and quality based upon service priority to ensure that critical communications are provided “assured access”. MUOS employs 120 levels of priority along with a set of congestion control mechanisms that are invoked as a function of service priority to manage communications resource consumption in the system. A pro-active philosophy is taken that incrementally degrades quality-of-service (QoS) (but still providing compliant QoS) with level of congestion to gracefully throttle resource consumption.*

## INTRODUCTION

MUOS requirements infer a simple pair of mechanisms to manage congestion in the system: priority-driven preemption and queuing. Such a view results in “reactive” behavior that is triggered by call events, making it operationally inefficient and requiring critical timing. In contrast, modern-day commercial systems have evolved to a “proactive” approach where communications resources are often shared with reliance on statistical multiplexing and resource utilization optimized via intelligent algorithms centered on real-time statistical analysis and forecasting methodologies. In this way, efficient and dynamic allocation and re-allocation of communications resources are enabled.

A major difference between military and commercial communication systems is that the former often demands dedicated resource allocations whereas the latter considers resources to be commodity that must be shared to the maximum extent possible to maximize return on investment of the infrastructure. Furthermore, while commercial system standards allow for the utilization of priority in allocating resources, vendors have been reticent to incorporate this capability into fielded systems. MUOS marries the unique needs of military users with methodologies proven effective by commercial systems in the area of resource utilization, resulting in the

Asynchronous, Aggregate Resource Management (AARM) model. AARM is comprised of several congestion control mechanisms up to and including preemption that are invoked as a function of priority level to gracefully manage communications resources to assure that high-priority communications are served immediately.

## BACKGROUND

Before describing how AARM manages communication resources, it is necessary to describe the basic architectural framework of MUOS and the services that consume the communications resources.

### A. MUOS Communications Architecture

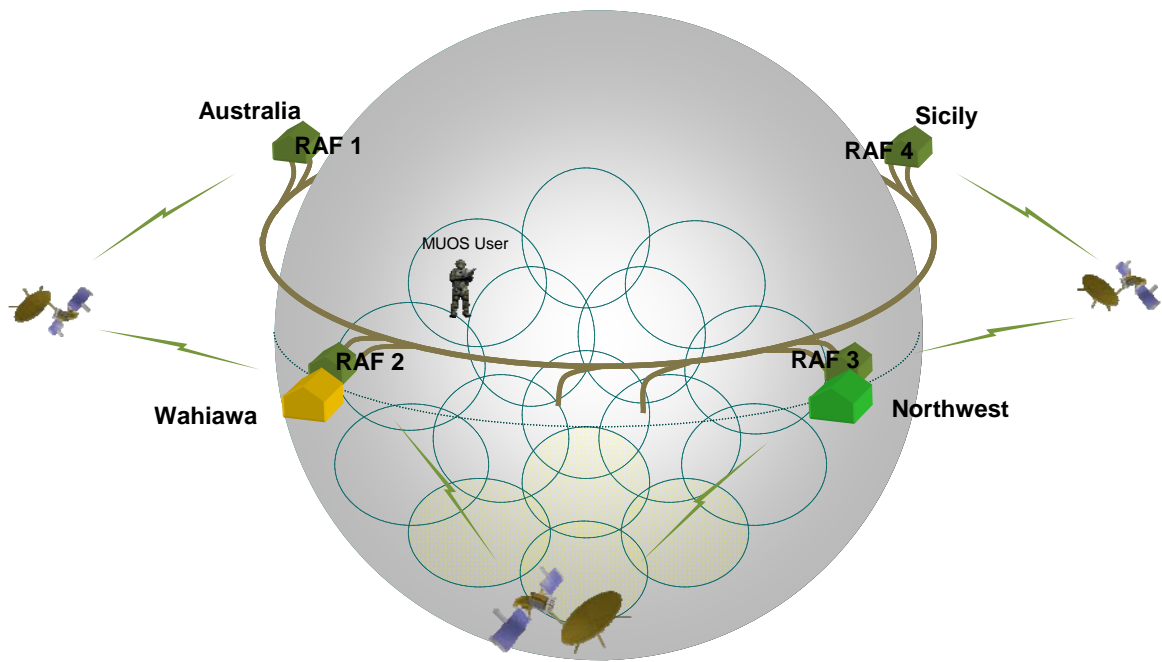
The MUOS communications architecture is based upon commercial Packet-Switched (PS) 3G WCDMA technology. Figure 1 illustrates the high-level physical topology of the MUOS network. Each of four operational geosynchronous satellites acts as a 16-beam “cell tower”. Each satellite beam can contain up to four WCDMA carriers (but at UHF frequencies). A “cell” is defined as a satellite beam and the corresponding carrier frequency. Each of four geographically spread Radio Access Facilities (RAFs) houses both the Radio Access Network (RAN) and the bearer traffic related portion of the Core Network (CN) equipment. Each RAF serves two satellites and each satellite serves as a cell tower for two RAFs. Two Switching Facilities (SFs) house the remaining portion of the Core Network equipment (e.g., authentication and subscriber database functions).

### B. MUOS Services

There are two types of services provided by MUOS.

#### 1) Point-to-Point Service

MUOS Point-to-Point (P2P) service is nearly identical to the services offered by commercial 3G WCDMA networks. UEs read broadcast System Information Blocks (SIBs) to obtain the requisite information to gain initial access to the network that subsequently allows them to first register with the network after which they can initiate service. A UE requesting service sends his request on the Random Access Channel (RACH) after which it is directed by the RAN with a response on the Forward Access Channel (FACH) to move to dedicated channel resources to complete the service setup control messaging. Once engaged in a



service, a UE consumes both a Dedicated Traffic Channel (DTCH) and a Dedicated Control Channel (DCCH). These logical channels can either map to Dedicated Channel (DCH) or the RACH / FACH resources depending upon the amount of traffic flowing (low levels of traffic are sometimes sent on common channels). P2P communications are unplanned, meaning that any registered UE can initiate a call to any other registered UE at any time provided sufficient communications resources are available.

## 2) Group Service

Group service is a significant departure from P2P service. The topology of a given group is pre-planned (although a mechanism exists to add additional cells “on the fly” if the group is operating outside the pre-planned topology) and can span many cells. An “enabled” group is a group that has been provisioned into the Group Manager in the Radio Access Network (RAN). An “active” group is an enabled group for which traffic is flowing. In Group service, user-to-base (U2B) and base-to-user (B2U) traffic channels are decoupled. Only one user “has the floor” at a time. The “talking” UE uses the U2B channel while all users in a cell share the same B2U channel to listen. Thus, at any time an active group consists of one and only one U2B link and as many B2U links as there are cells defined for the group. Group broadcast messaging carried on a Forward Access Channel (FACH) informs User Equipment (UEs) of which groups are enabled and the salient parameters needed to initiate/participate in each (carrier frequency for example). A UE that wishes to “activate” a specific enabled group sends a talk request on the Random Access Channel

(RACH) and then immediately begins to transmit bearer data on a U2B Dedicated Traffic Channel (DTCH) whose parameters are implicitly determined based on information that includes the unique ID of that group along with the satellite and beam. If the request is granted, the dedicated channel assignment information is sent in the grant message on the FACH to inform listening UEs of where to “tune” to hear the bearer traffic being transmitted for the group. The grant message is sent on all cells for which the group is defined. When the talking UE is done, it sends a release request to which the RAN responds with a release indication on the Dedicated Control Channel (DCCH). The request – grant - release procedure takes place every time there is a change in talker and facilitates “Group agility”, meaning that a different type of service can be requested by each talker; e.g., a voice burst initiated by one group member can be followed by a data burst by another.

## BODY

Now that a general description of MUOS has been put forward, we describe the AARM approach to managing communications resources.

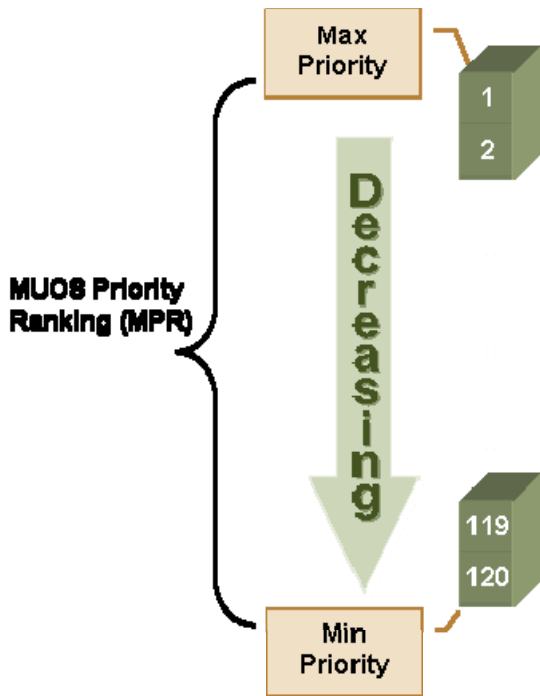


Figure 2

### A. Priority Scheme

In order to adjudicate communications resources by priority, it is necessary to provide a priority scheme. The MUOS priority scheme is governed by the 120-level MUOS Priority Ranking (MPR). The maximum priority corresponds to an MPR value of 1 while the minimum priority is designated as MPR = 120.

There are 108 priorities reserved for Group communications. Each group is provisioned by planners with a MUOS Priority Designator Index (MPDI) that maps one-to-one with MPR. An MPDI-to-MPR mapping table resides in the RAN. Planners can alter the relative priorities simply by altering the MPDI-to-MPR mapping table. UEs are not cognizant of the priority of a group, but the Group Manager in the RAN is.

For P2P services, the priority of a service is established at call setup by the requesting UE (as selected by the user). Because of the requirement to be compatible with DSN Multi-level Precedence and Preemption (MLPP), a P2P service assumes one of the following MLPP precedence values: Routine, Immediate, Priority, Flash, Flash Override. The MLPP precedence values map to MPR values. The mapping is defined by planners with the mapping table residing in the RAN. Because of the desire to be able to differentiate between voice and data for P2P communications, there are actually a total of 10 P2P priority values (five for voice and five for data). The 10 P2P priorities combined with the aforementioned 108

Group priorities leave 2 spare MPR values.

### B. Communications Resources

Table 1 – Key communications resources

Resource	Description
UL Interference	Multiple access interference MAI in the U2B
DL Carrier Power	Aggregate carrier power in the B2U
UL ASEs	Resource utilization in the U2B normalized to that consumed by speech call.
DL ASEs	Resource utilization in the B2U normalized to that consumed by speech call.
UL CEs	Measure of radio base station (RBS) resource consumption in U2B.
DL CEs	Measure of radio base station (RBS) resource consumption in B2U.
DL Channelization Codes	Total B2U channelization codes consumed.
RACH	RACH channel utilization.
FACH1	FACH1 channel utilization.

Table 1 presents the key communications resources on a cell basis. It is the consumption of these resources that AARM manages through a series of congestion control mechanisms (described in the next section).

### C. Congestion Control Mechanisms

Table 2 presents our congestion control mechanisms and their effects on resources in the U2B and B2U directions. Each mechanism is explained in more detail in what follows. A key attribute of our approach is that, when appropriate, the UE is part of the control mechanism, facilitated by including the congestion state of the system in SIB information (P2P) and group broadcast information (Group). The state indicates, based on MPR, when services are subject to the various control mechanisms.

Table 2 - Congestion Control Mechanisms

Action	U2B	B2U
Enable DTI	yes	no
Channel Switching to CELL_FACH	no	yes
No operation above QBR UL	yes	no
No operation above QBR DL	no	yes
Reject/preempt multi-service requests	yes	yes
Reject/preempt new service requests	yes	yes
Access Service Classes	yes	no
Group NAK Control	yes	no

## *Dedicated Channel Resources*

### *a) DTI*

Dovetail Interleaving (DTI) improves capacity by extending the transmission time of an FEC coded block in order to enhance the time diversity of the link. Time diversity can reduce the  $E_b/N_0$  required to close a link, which in turn increases the capacity of a WCDMA air interface. When there is no congestion, there is no need to invoke DTI. When UL interference begins to enter an “uncomfortable” range, however, it is best for the system to have DTI enabled for some or all of those services that can tolerate the latency that it introduces (everything except voice and streaming data). DTI is never invoked mid-call; it is invoked as part of call setup. In P2P communications, the UE compares the precedence of a new (DTI eligible) service request against the precedence at which DTI is to be invoked and configures the service accordingly. For Group communications, the Group Manager determines which groups are subject to DTI based on the congestion state and updates the group broadcast messaging accordingly.

### *b) Channel Switching Between Dedicated and Common Channels*

As stated previously, bearer traffic is sometimes mapped to the common channels (RACH/FACH) when the traffic does not warrant dedicated channel resources. We treat the channel-switching between common and dedicated channel resources as a congestion control mechanism. When there are plenty of DL channelization codes available, a UE is maintained on dedicated channels. As the number of available DL channelization codes in a beam-carrier begins to dwindle, channel-switching to common channels is enabled, as a function of precedence level. This mechanism is transparent to the UE in that it simply follows the channel-switching directions given by the RAN.

### *c) Operation Limited to Quality of Service Bit Rate (QBR)*

Associated with each service is the Quality of Service (QoS) Bit Rate (QBR). This is the minimum acceptable rate for the service. Services also have associated with them a Maximum Bit Rate (MBR) which can be thought of as the upper limit that the service will run. For streaming and voice services,  $MBR = QBR$ . For other data services, however, typically  $MBR > QBR$ . Limiting UEs to operating at their QBR can help both U2B and B2U. In U2B, it decreases power and MAI. In B2U it decreases power and downlink channelization code usage.

In P2P communications, the channel can be switched real-time between QBR and MBR in accordance with the

bandwidth demand of the service. In the B2U (DL) direction, the QBR limit is executed by the RAN performing B2U channel switching (P2P communications only). In the U2B (UL) direction, the QBR limit is executed by the UE policing itself, only allowing transmissions at the QBR. For P2P communications, it is possible to have a QBR limit in only the UL, only the DL, or both. The UE is aware of whether is subject to QBR limiting based upon broadcast information in a SIB.

In Group communications, there is no channel-switching in the middle of a talk burst. A QBR limit, however, dictates what a UE that wishes to talk can request. Also, because of the tight coupling between UL and DL, if there is a QBR limit in either the UL or DL in any cell for which a given group is defined, then the limit must exist in ALL cells in both the UL and DL, which effectively means we need only limit it in the UL (since it is the talk request on the uplink that defines the resources assigned on the shared DL). Stated more succinctly, any QBR limit in either the UL or DL in any beam-carrier of a Group must result in a UL QBR limit in all beam-carriers defined for that group. The Group Manager indicates which groups are subject to QBR limits in group broadcast messaging based upon the information it receives from the Congestion Manager.

### *d) Reject/Preempt Multiple Simultaneous Services*

Another way to reduce congestion is to limit simultaneous services. Because Group does not support simultaneous services (instead Group has agility to setup a different service for each PTT), this affects P2P service only. The multi-RABs defined to implement simultaneous services were not designed for maximum radio resource efficiency particularly concerning downlink channelization codes. So, the ability to prevent a UE from using multi-RABs by precluding simultaneous services is desirable.

We reached the conclusion that preemption and blocking of new service requests would occur at the same level of precedence. To be consistent, we believe that if we are blocking new request for multi-services at priority X, then we should be preempting one side of existing multi-services at the same level. Based upon information in a SIB, a UE knows whether it is eligible to engage in multiple simultaneous services or not (based upon the precedence of the services) and behaves accordingly. When not eligible for multiple services, the UE would not allow initiation of a 2<sup>nd</sup> service or would release one of two services if the notification arrived after two were already established.

e) *Reject/Preempt Single Services*

Based upon the congestion state of the system, new requests for service can be rejected and existing services can be preempted.

In P2P communications, the UE compares the precedence level of a new service request to the precedence level at which new service requests are blocked by the network. If the new request is of insufficient precedence, then the request is not sent OTA to the RAN; the UE quells it. When a request is rejected on the originating side, the originating UE queues the request. When the call is rejected on the terminating end, the originating UE does not queue the request. If the UE has an on-going service, it compares the precedence of that service against the precedence at which services are subject to preemption. If the service is indeed subject to preemption, the UE initiates release of the service.

In Group communications, rejecting requests for new service simply translates to suppressing the activation of a group in the congested cell. The Group Manager determines which groups are ineligible for activation based upon the congestion information it receives from the Congestion Manager. The UE determines if its group is subject to having activations suppressed via information contained in broadcast messaging as configured by the Group Manager. Preemption is carried out by the Group Manager. The Congestion Manager sends a message to the Group Manager containing the congestion status information which includes the precedence level at which preemption is to be invoked. The Group Manager determines which groups are subject to preemption and configures group broadcast messaging accordingly. For each group to be preempted for which traffic is flowing, it sends a release message on the DCCH of all cells defined for that group. For any group to be preempted for which there is no traffic, suppressing its activation is sufficient. It should be noted that while a QBR limit in any cell of a group results in a QBR limit in all cells for that group, the same is not true for preemption. Preemption of a group occurs on a cell by cell basis and results in a partial group if there are any remaining cells defined for the group for which communications are allowed.

To this point, we have tacitly assumed that the decision to preempt/block is based on both UL and DL resources. For P2P, such an assumption presents no real limitation since the ability to both “speak” and “listen” are necessary conditions of a P2P service. For Group, however, since the UL and DL are independent, the inhibition of one should not necessarily inhibit the other. Specifically, if preemption/blocking is invoked in a cell (at a given MPR)

due to congestion on the UL, we may still want to send replicated traffic (which would have originated on a different cell) on the DL (i.e., the Group Manager would replicate group traffic on the DL of a group whose activation is suppressed on the uplink). Clearly, such a capability provides benefit only in the case of Groups that span more than one cell and that are being inhibited by congestion in the UL only (due to the Group having insufficient precedence). If preemption/blocking is invoked in a cell (at a given MPR) due to congestion on the DL, however, we do wish to prohibit the UL in that beam-carrier as well since a UE that is the “talker” is relying on dedicated channel information on the DL for control plane functionality (e.g., group power control).

*Common Channel Resources*

To this point, discussion has centered on dedicated channel resources. Management of common channel resources is also needed, however. Analysis of system behavior indicates that the preemption of services is sufficient to control traffic on FACH, so no special mechanism is required to address it. We have identified, however, the need to include additional control mechanisms to address RACH.

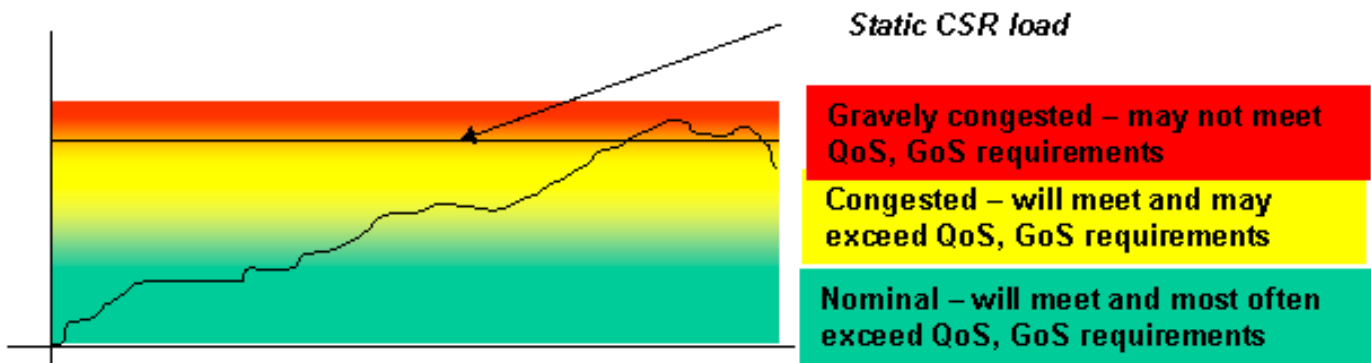
By including functionality to quell both P2P service requests and group talk requests, we have provided a powerful mechanism to help prevent congestion on the RACH due to P2P traffic mapping to CCCH and Group traffic mapping to GCCCH. It does not address P2P traffic mapping to DTCH or DCCH nor Group traffic on DTCH which carries the NAKs when RLC AM is used. Hence, we introduce two additional mechanisms.

1) *Access Service Classes*

Access Service Classes (ASCs) are a standard UMTS method to help control traffic on RACH. With each ASC there is an associated persistence value. The UE generates a random number prior to transmitting on the RACH and compares it to the persistence value of the associated class. Depending upon the result, the UE is either allowed or not allowed to transmit and must wait until the next transmission interval and try again. We invoke this capability on MUOS to help throttle traffic on RACH.

2) *Group NAK Control*

Some data service requirements results in configuration of Layer 2 acknowledged behavior. In Group communications, NAKs (No AcKnowledgements) are sent on the RACH. Because there are potentially many UEs in



**Figure 3 – Operation philosophy**

a group, the volume of NAK traffic could consume a large portion of the RACH channel. Therefore, we define a means to allow a UE to accumulate multiple lost PDUs before sending a NAK to request their retransmission. In this way, a single message can be sent to effectively address a multiplicity of NAKs. For each group, there is a provisioned parameter defining how long to wait prior to sending an RLC NAK. The Congestion Manager determines at what precedence level this function is invoked and passes this information as part of the congestion status indication to the Group Manager. The Group Manager then determines for which groups this function applies. Thereafter, for any group for which it applies, the Group Manager sends the wait time in the message that grants service.

**D. Operation**

Figure 3 presents a notional depiction of AARM operation in response to traffic demand/resource consumption. The control mechanisms defined previously are invoked with increasing degree as more communications resources are consumed. At low levels, there is no need to invoke congestion control mechanisms. When this is the case, services will experience maximum data rates and minimum latencies. As more resources are consumed, the control mechanisms begin to be turned on at low levels. For the dedicated channel congestion control mechanisms (except single service blocking/preemption) this corresponds to invoking them at high values of MPR (recall, low priority corresponds to large MPR values). As more resources are consumed, these mechanisms are invoked at lower values of MPR until at an MPR of 1, all services are subject to DTI, channel switching to common channel, QBR limits, and restrictions on multiple services. This point corresponds to the required capacity of the system as any single service can meet its QoS and GoS requirements with the aforementioned control mechanisms invoked. Above this level system capacity is exceeded meaning that QoS and GoS may not be met, thus necessitating that

preemption of single services take place (again as a function of MPR).

This approach attempts to minimize the need to preempt services by taking proactive actions to reduce resource consumption by other means during times of congestion while allowing services to operate without congestion restraints when resources plentiful.

**SUMMARY**

We have presented a means to control the consumption of communications resources in MUOS that incrementally invokes a suite of congestion control mechanisms based upon resource consumption and service priority to ensure that high priority services always receive the best service available. This methodology allows services to experience QoS and GoS that is better than required when the network is not congested while preserving QoS and GoS as congestion grows. When demand for resources exceeds required system capacity, preemption is invoked based on MPR to ensure that high priority services are assured access.