What are Microbursts?

Microbursts are traffic patterns where traffic arrives in small bursts. While almost all network traffic is bursty to some extent, storage traffic usually has large bursts when transferring data. Applications in financial trading environments or in Web2.0 environments with Memcached servers can have very small burst sizes. Such microbursts usually last only a few microseconds and are hard to detect using standard network management tools that sample data rates over a few minutes.

When only one server is communicating with another server in the network, bursts are usually not a problem as there is no oversubscription. However, when many servers (initiators) are talking to one target then the bursts can result in fan-in. Fan-in due to microbursts can cause short-term packet loss and sufficient buffering is needed on the networking interconnect to handle these scenarios.

Myths about Microbursts

The common myths about microbursts are:

- These are special traffic patterns that create more stress on a switch
- Low latency cut-through switches do not have the buffering needed to handle microbursts
- Microbursts result in higher latency and jitter

One reason these myths exist is that legacy switches could not handle such traffic patterns. In the legacy designs, sufficient packet memory bandwidth was not allocated and a bursty pattern with fan-in could use up all of the memory bandwidth inside the switch. However, ultra-low latency switches can now support close to a terabit of bandwidth and uniform performance is observed with all traffic patterns.

Cut-through switches, such as the Arista 7100 series, can forward all traffic patterns with no impact on latency. This applies to both unicast and multicast traffic. However, under a fan-in scenario where “N” producers are sending to one consumer, packets will be stored in packet memory. Latency will obviously be higher in the fan-in case as packets await their turn in transmit-queues. This is true for all congestion scenarios and has nothing to do with microbursts.

Impact on Application Performance

Applications rely on the network to provide a fabric for inter-node communication. Any drops in the network can result in poor performance at the application level. Hence, it is important to have sufficient buffering in the switches to handle such congestion.. When the network can buffer bursts, whether they are microbursts or normal bursts, applications can achieve optimal performance.

Fan-in results in a congestion point and queue build up and the end-to-end latency increases in such scenarios. While higher latency is never preferred, the increased latency due to queue delays has a much lower impact on application performance compared to any packet loss. When the switches can buffer traffic, the applications do not time-out and the TCP window can scale to its maximum size.

Arista offers two approaches to address the buffering needs on switches:

- 7100 Series: Ultra-low latency, 1/10GbE cut-through switches
- 7048 & 7500 series: Low latency, deep buffer, store-and-forward switches

The ultra-low latency 7100 series switches are ideal for environments that require the lowest latency. The switches have enough buffering to handle bursts and fan-in for most real world application environments.

The 7048 switch is best suited for speed mismatch (1GbE node communicating with 10GbE node) or sustained congestion scenarios. The deep buffer architecture is optimal for storage, video, or more demanding 1GbE connected nodes. The 7500 switches have some of the deepest buffers to address some of the most extreme cases of congestion.

Testing for Microbursts & Fan-In

Any test to simulate real world traffic needs to take into account what the actual traffic patterns are for a given application. Attached is a report generated using Ixia, a traffic generator to simulate real world traffic. The network scenario being simulated is shown below.

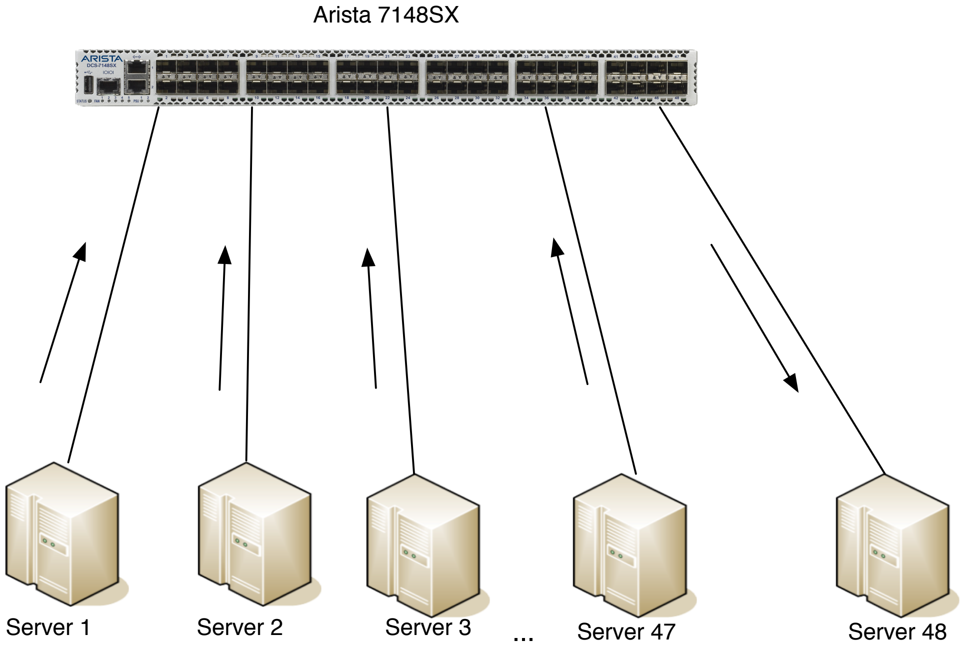

Figure: 47 servers (initiators) sending bursts to one target

In this setup, Servers 1 through 47 (initiators) are simultaneously sending bursty traffic to Server 48 (target). The traffic pattern consists of 256 byte packets with a burst size for each stream of 30 packets. Thus, Server 48 (the target) receives 1,410 packets at the same time. Average packet size of 256 bytes fairly represents the small but bursty messages in financial messaging applications or in Memcached lookups. This IXIA test results can be viewed at the end of this report. No drops were seen and the lossless transmission in this microburst environment highlights the high performance of the Arista 7100 series switches.

Conclusion

Understanding the exact traffic pattern for any application is important to analyze overall performance. Microbursts are just another traffic pattern though they are harder to detect due to their transient nature. Modern switches can handle such bursts well. Fan-in can result in congestion and sufficient buffering is needed on the switches to handle these scenarios.

The Arista 7100 series switches offer ultra-low latency of 600ns (7124S) to 1.2us (7148SX) and provide a lossless fabric with a unique cut-through architecture and buffers to handle congestion. These switches are ideal for 1/10GbE deployments, including financial trading applications, video, storage, or web environments.

Overview: Cisco commissioned Miercom to compare Arista 7100S series switches against the Cisco Nexus 5000 series switches. The tests used so called “typical real world application traffic scenarios” and were focused on finding thresholds where packets were dropped by the respective switches. To Miercom's credit, these tests are clearly described, correctly run, and accurately reported.

Arista objects to Cisco's characterization of these tests as being similar to "real world" traffic. This claim is deeply misleading. These tests were clearly designed to make the Nexus 5010 appear better than the Arista 7124S at all costs. **Most of the testing was focused on fan-in with multicast cases – as much as 23-to-1 or even 47-to-1 fan-in.** Each test case is custom – no standard RFCs were used, which have been a benchmark in the industry for several years now to compare products.

Product Architectures: First, it is important to understand the product architectures – we will leave it to the reader to decide the co-relation between the architecture and the tests noted in the report.

Nexus 5000: The Nexus 5000 buffer memory is statically associated with a given input port, 480KB per port for each of its ports. Nexus buffer is allocated in chunks of 140 bytes.
Arista 7100: The 7124S has 2MB of centralized packet buffer. After subtracting off various overheads, there are 1718KB of usable memory with 20KB reserved per port and an additional 1238KB usable by any port experiencing congestion. The 7148SX has 4 times the memory of 7124S.

The two architectures are substantially different when it comes to buffering. The Nexus 5000 will perform better when most ports are congested at the same time. Arista 7100 switches will perform better when some ports get congested at any given time, which is the common case in most networks.

What was Tested: 96 and 128 byte packets were used. Tests were n-to-many unicast or multicast – essentially the same test repeated over and over again where as many as 23 or 47 sources were used, all simultaneously sending packets to each of the remaining ports, one-by-one. This ganging up on the destination port results in massive congestion, resulting in packet drops. Multicasts, as we all know it is a one-to-many protocol. Not many-to-one! In these tests, that was turned around, simply to cause congestion and show drops. Which network in the world has 23 hosts sending to one destination, at exactly the same time?

What happens in the real world: In real world financial trading networks, multicast feeds come in from the exchanges (such as NASDAQ, NYSE, BATS, etc.). These multicast feeds are processed for synchronization, accuracy and optimizations before being relayed to algorithmic trading engines. In today’s trading networks, all available market data feeds added together sum to 3.6Gbps (obtained by adding all feeds available through Savvis). Thus, the massive many-to-many fan-in cited in Cisco’s report is not seen in real world networks.

Why 96 and 128 byte packets? Nexus buffer is allocated in chunks of 140 bytes. Thus by picking packet sizes that fit nicely in the Nexus architecture, Cisco was able to achieve better performance.

What happens in the real world: In real world financial networks, the packets are of mixed sizes. The NYSE market data feed for example contains packets from 64 bytes to 1273 bytes. The average packet size is 192 bytes. Thus the tests skipped a key component: using real world traffic. In addition, since various packet sizes appear in data feeds, a latency test with random packet sizes, or an IMIX pattern needs to be used. With such tests, Nexus 5000 switches have much higher jitter (3 to 4 usec), while the Arista 7100 series switches can handle such patterns with less than 30 nanoseconds of jitter.

Congestion & Latency:

In all of these custom tests, the objective was to find the breaking point of the Arista 7100S switches. However, the most common cases of congestion, 2:1 fan in, was not tested. In this most relevant, real-world case, the Arista 7100 switch outperforms the Nexus 5000.

The test summary argues that low latency is not a good way to measure performance if you drop packets during congestion. Well, that depends on the application and how much congestion needs to be handled before dropping a packet. In some applications, low latency is critical - buffering traffic for 50 usec or more does not help. The solution in these low-latency cases is to identify the sources of congestion and eliminate them by re-architecting the network.

Arista does understand the need for buffering in various applications. In cases where more buffering is required, sub-micro second latency is not a requirement - such as storage. Arista 7048 and Arista 7500 series switches offer very deep-buffer solutions for both 1GbE and 10GbE, buffering over 40ms of traffic on all ports simultaneously.

Summary

- All tests commissioned by Cisco have the same basic structure. Each sender sweeps across all recipients in perfect lock-step, simultaneously blasting each recipient with line-rate bursts of small packets, so as to create uniform congestion across all ports simultaneously, hiding the Nexus' inflexible buffer allocation policy while exaggerating the impact of a shared memory buffer.
- These tests use exclusively packets that are just under 140 bytes in size. At this size, the Nexus 5010 uses 140 bytes of buffer per packet (wasting a tiny handful), whereas the 7100 uses 512 bytes of buffer per packet (wasting 384 bytes of buffer per packet)**.**
- The tests spread congestion uniformly across all ports of the switch. Real world scenarios have anything but uniformly spread congestion.
- While any-to-any multicast exists in financial trading applications, with all market data feeds combined at 3.6Gbps today, these “micro-bursts” are no longer an issue in today’s low-latency 10GbE networks.

Conclusion

Custom test reports can make one product look better than the other. All customers are encouraged to do their own analysis before making a purchasing decision.

While the results of these tests are valid, the tests themselves do not resemble anything in the real world. Instead, these tests commissioned by Cisco are designed to find extreme corner cases to show the Nexus 5010 performing better than the Arista 7124S while hiding the deficiencies of the Nexus 5000 product line. An industry wide comparison was done by Network World recently where all vendors were invited to participate under a common test criterion. Arista 7100 switches outperformed all other switches in that testing, including Cisco’s Nexus 5000.

Arista has over 100 customers in low latency financial applications – customers who are happy with the performance of our products and don’t see any of the cases cited in the report.

For more information, please contact info@aristanetworks.com or visit www.aristanetworks.com.

IxNetwork Report

Test: **Custom Microburst Test**
Test Date: **09/29/2009** **20:43:06**

www.ixiacom.com
www.ixiacom.com/support

Traffic Item Statistics

Traffic Item	Tx Frames	Rx Frames	Frames Delta	Loss %	Tx Frame Rate	Rx Frame Rate	Rx Bytes	Rx Rate (Bps)	Rx Rate (bps)	Rx Rate (Kbps)	Rx Rate (Mbps)	First TimeStamp	Last TimeStamp	Dead Flow
MicroBurst	1410	1410	0	0	0	0	360960	0	0	0	0	00:01.5	00:01.5	0

User Defined Statistics

Source Port	Tx Frames	Rx Frames	Frames Delta	Loss %	Tx Frame Rate	Rx Frame Rate	Rx Bytes	Rx Rate (Bps)	Rx Rate (bps)	Rx Rate (Kbps)	Rx Rate (Mbps)	First TimeStamp	Last TimeStamp
Port01	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port02	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port03	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port04	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port05	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port06	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port07	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port08	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port09	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port10	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port11	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port12	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port13	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port14	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port15	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port16	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port17	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port18	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port19	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port20	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port21	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port22	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port23	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port24	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port25	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port26	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port27	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port28	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port29	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port30	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port31	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port32	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port33	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port34	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port35	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port36	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port37	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port38	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port39	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port40	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port41	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port42	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port43	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port44	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port45	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port46	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port47	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5

Data Plane Port Statistics

Port	Tx Frames	Rx Frames	Tx Frame Rate	Rx Frame Rate	Rx Bytes	Rx Rate (Bps)	Rx Rate (bps)	Rx Rate (Kbps)	Rx Rate (Mbps)	First TimeStamp	Last TimeStamp
Port01	30		0								
Port02	30		0								
Port03	30		0								
Port04	30		0								
Port05	30		0								
Port06	30		0								
Port07	30		0								
Port08	30		0								
Port09	30		0								
Port10	30		0								
Port11	30		0								
Port12	30		0								
Port13	30		0								
Port14	30		0								
Port15	30		0								
Port16	30		0								
Port17	30		0								
Port18	30		0								
Port19	30		0								
Port20	30		0								
Port21	30		0								
Port22	30		0								
Port23	30		0								
Port24	30		0								
Port25	30		0								
Port26	30		0								
Port27	30		0								
Port28	30		0								
Port29	30		0								
Port30	30		0								
Port31	30		0								
Port32	30		0								
Port33	30		0								
Port34	30		0								
Port35	30		0								
Port36	30		0								
Port37	30		0								
Port38	30		0								
Port39	30		0								
Port40	30		0								
Port41	30		0								
Port42	30		0								
Port43	30		0								
Port44	30		0								
Port45	30		0								
Port46	30		0								
Port47	30		0								
Port48		1410		0	360960	0	0	0	0	00:01.5	00:01.5

Flow Statistics

Tx Port	Rx Port	Traffic Item	Source Port	Tx Frames	Rx Frames	Frames Delta	Loss %	Tx Frame Rate	Rx Frame Rate	Rx Bytes	Rx Rate (Bps)	Rx Rate (bps)	Rx Rate (Kbps)	Rx Rate (Mbps)	First TimeStamp	Last TimeStamp
Port01	Port48	MicroBurst	Port01	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port02	Port48	MicroBurst	Port02	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port03	Port48	MicroBurst	Port03	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port04	Port48	MicroBurst	Port04	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port05	Port48	MicroBurst	Port05	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port06	Port48	MicroBurst	Port06	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port07	Port48	MicroBurst	Port07	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port08	Port48	MicroBurst	Port08	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port09	Port48	MicroBurst	Port09	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port10	Port48	MicroBurst	Port10	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port11	Port48	MicroBurst	Port11	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port12	Port48	MicroBurst	Port12	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port13	Port48	MicroBurst	Port13	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port14	Port48	MicroBurst	Port14	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port15	Port48	MicroBurst	Port15	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port16	Port48	MicroBurst	Port16	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port17	Port48	MicroBurst	Port17	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port18	Port48	MicroBurst	Port18	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port19	Port48	MicroBurst	Port19	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port20	Port48	MicroBurst	Port20	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port21	Port48	MicroBurst	Port21	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port22	Port48	MicroBurst	Port22	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port23	Port48	MicroBurst	Port23	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port24	Port48	MicroBurst	Port24	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port25	Port48	MicroBurst	Port25	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port26	Port48	MicroBurst	Port26	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port27	Port48	MicroBurst	Port27	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port28	Port48	MicroBurst	Port28	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port29	Port48	MicroBurst	Port29	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port30	Port48	MicroBurst	Port30	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port31	Port48	MicroBurst	Port31	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port32	Port48	MicroBurst	Port32	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port33	Port48	MicroBurst	Port33	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port34	Port48	MicroBurst	Port34	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port35	Port48	MicroBurst	Port35	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port36	Port48	MicroBurst	Port36	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port37	Port48	MicroBurst	Port37	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port38	Port48	MicroBurst	Port38	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port39	Port48	MicroBurst	Port39	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port40	Port48	MicroBurst	Port40	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port41	Port48	MicroBurst	Port41	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port42	Port48	MicroBurst	Port42	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port43	Port48	MicroBurst	Port43	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port44	Port48	MicroBurst	Port44	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port45	Port48	MicroBurst	Port45	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port46	Port48	MicroBurst	Port46	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5
Port47	Port48	MicroBurst	Port47	30	30	0	0	0	0	7680	0	0	0	0	00:01.5	00:01.5