Green Flash

High performance computing for real-time science

WP7: Middleware study and down-selection

Final Prototyping Review, Paris 6/4/2018

Project #671662 funded by European Commission under program H2020-EU.1.2.2 coordinated in H2020-FETHPC-2014

Middleware study and down-select

- Study of existing middleware
- Identify candidate technologies:
  - Review the field and shortlist
- Down-select:
  - Assess against GF performance requirements and non-functional criteria

Middleware domains

- Data Pipeline
  - Temporal determinism
  - High bandwidth
  - Point – to – point
- Command/Control
  - RPC paradigm
  - Synchronous and asynchronous
  - Low bandwidth
- Telemetry
  - Publish/subscribe
  - High bandwidth

Common Criteria

ID	Criterion	Description	Weighting
DS-G-1	Cost	Lower cost or free solutions are preferable	1
DS-G-2	Ease of use	Middleware should support custom application development with minimum complexity and resources.	1
DS-G-3	Long-term support	Support for the technology should be anticipated from a supplier or other source for 10+ years into the future.	2
DS-G-4	Standards compliance	Standards ratified by a standards body are to prefered to de-facto standards. This is related to DS-G-3 in that standards are usually supported or implemented by multiple sources	2
DS-G-5	Familiarity	Technologies for which expertise already exists within the consortium/responsible partner(s) are to be preferred for prototyping.	2
DS-G-6	Commonality	Proliferation of technologies should be avoided, all else being equal, through the use of the same technology in multiple domains within Green Flash	1
DS-G-7	Power efficiency	This is an important goal for the Green Flash project and must be considered where appropriate as part of the down-selection process	2
DS-G-8	Source of Supply	Multiple sources of supply are to be preferred. This criterion is related to DS-G-3 and DS-G-4	2

Technology shortlist

Role	*Option 1*	*Option 2*
Pipeline	ZeroMQ	MPI
Command/Control	DDS	ICE
Telemetry	DDS	ZeroMQ

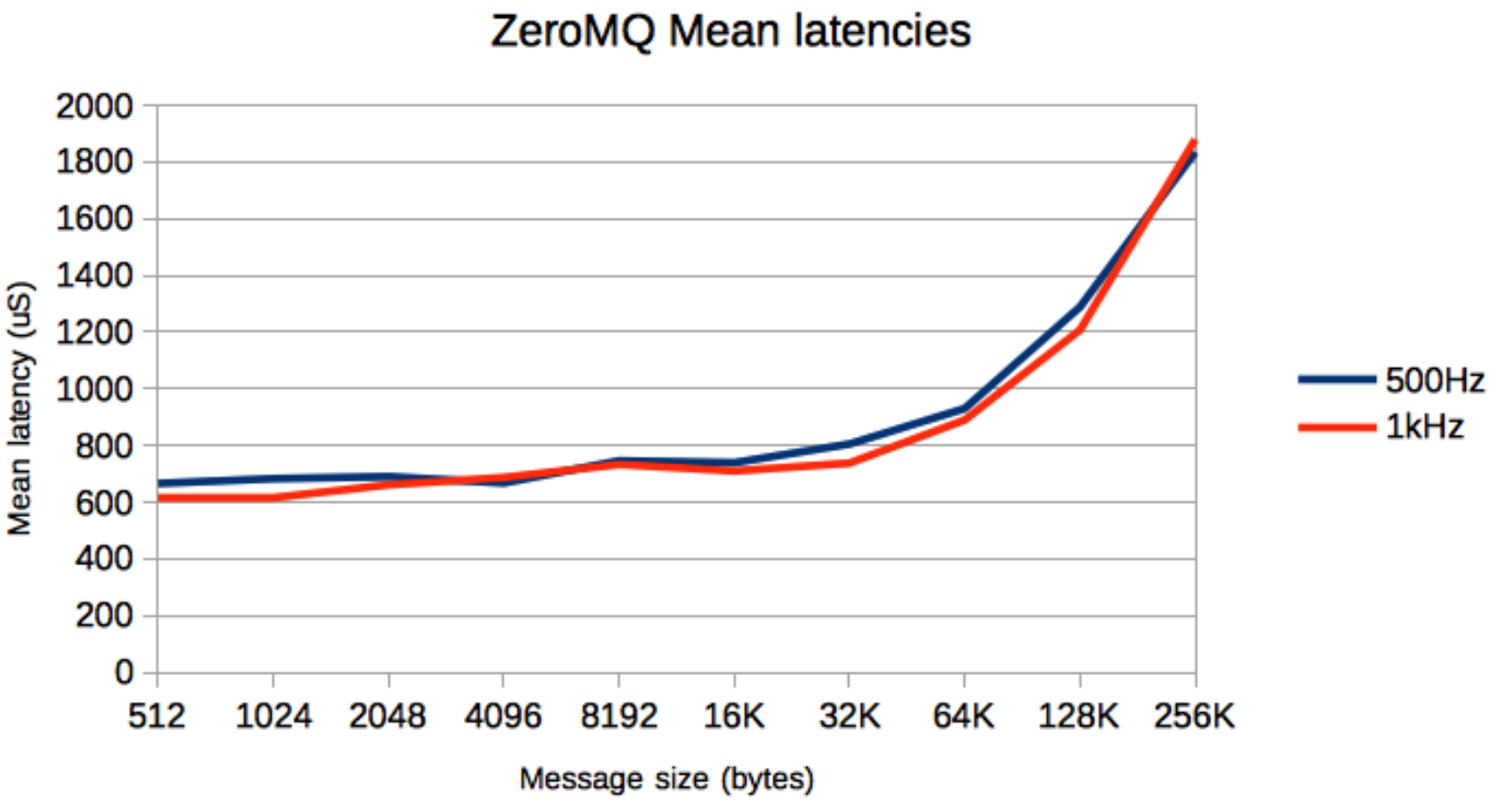
Criteria: real-time pipeline

ID	Criterion	Description	Weighting
DS-MW-1	Reliability	The middleware should be able to guarantee delivery of uncorrupted data, or at the least, detect and signal non-delivery or data corruption.	3
DS-MW-2	Latency	2mS (goal: 1mS) between first pixel received and last actuator demand delivered. This is the total latency budget for the pipeline, the majority of which must be available to be expended on processing; a nominal 10% of the budget has been allowed in this assessment for communications.	3
DS-MW-3	Jitter	100uS peak-to-peak; as in the case of latency, this is the budget for the pipeline. Contributions to jitter from different sources (processing, communication,...) sum quadratically; a nominal 30% of total jitter has been allowed.	3
DS-MW-4	Throughput	Within the pipeline: the most demanding case int terms of aggregate throughput is METIS LTAO mode, with a frame rate of 1kHz and 6 LGS/3 NGS WFS. The input bandwidth for pixel data is ~ 200Gb/s (25 Gb/s). However, this is not carried by a single connection, and pixel input data is not carried by the middleware. A more realistic requirement on bandwidth per link *within* the pipeline is the transport of pixel data for a single WFS, from a calibration module to a centroider module; for a single LGS WFS, the required bandwidth is 2.19 Gb/s (274 MB/s). If calibration and centroiding are perfomed within the same hardware module and data is not required to be transported on the network at pixel rates, the requirement is reduced to transporting frames of centroids from a single WFS, and for a LGS WFS at 1kHz this is evaluates to 350 Mb/s (44 MB/s).	3

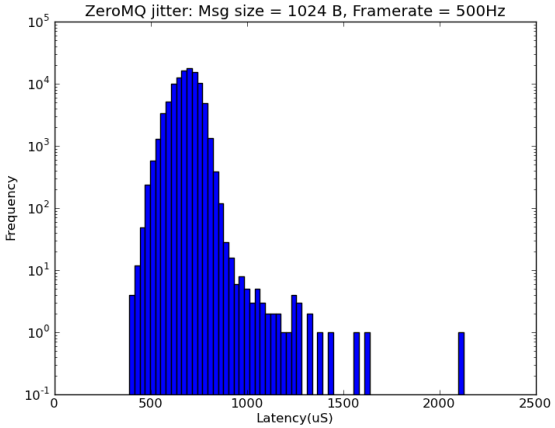
Test environment

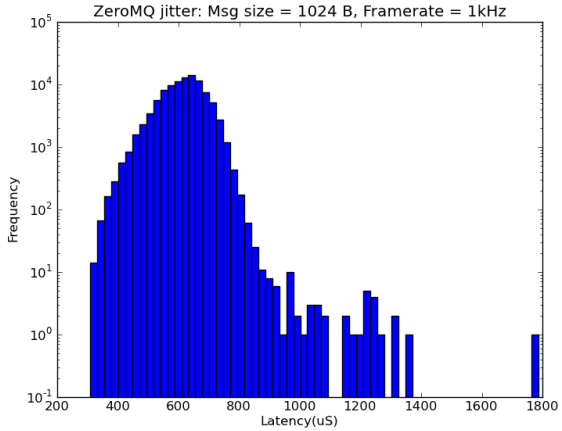
- Round-trip timing of messages: master → slave → master
- Range of message sizes, at frame rates of 500Hz and 1kHz
- 10Gbe ethernet, direct connection (no switch)

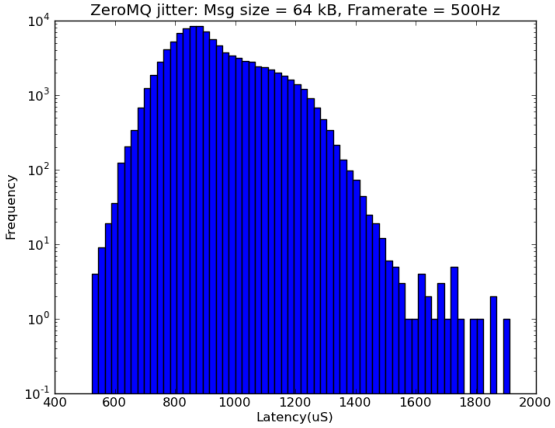
ZeroMQ: latency

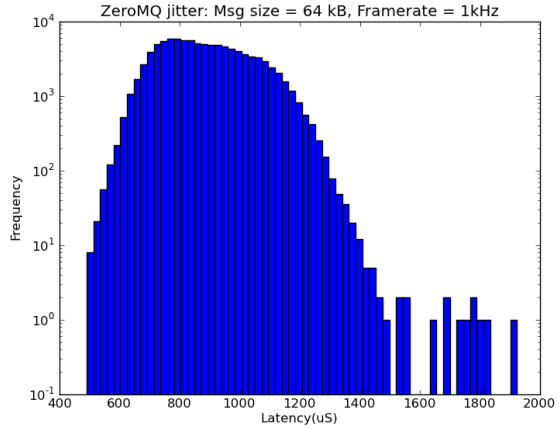


ZeroMQ: jitter



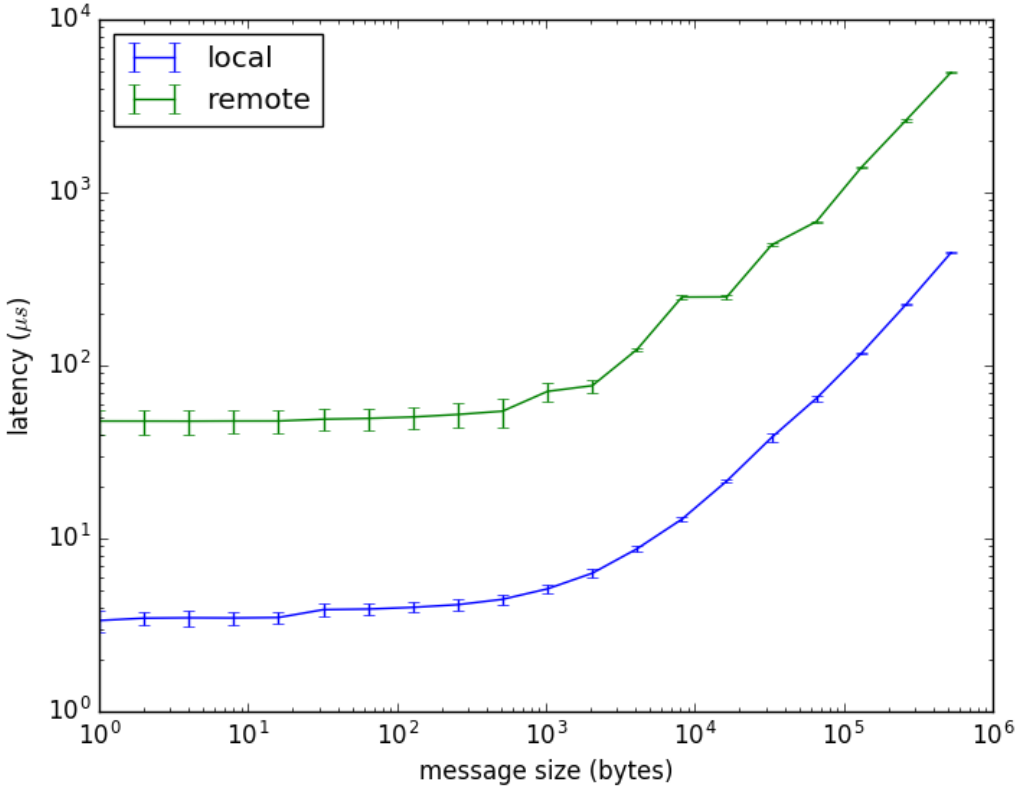






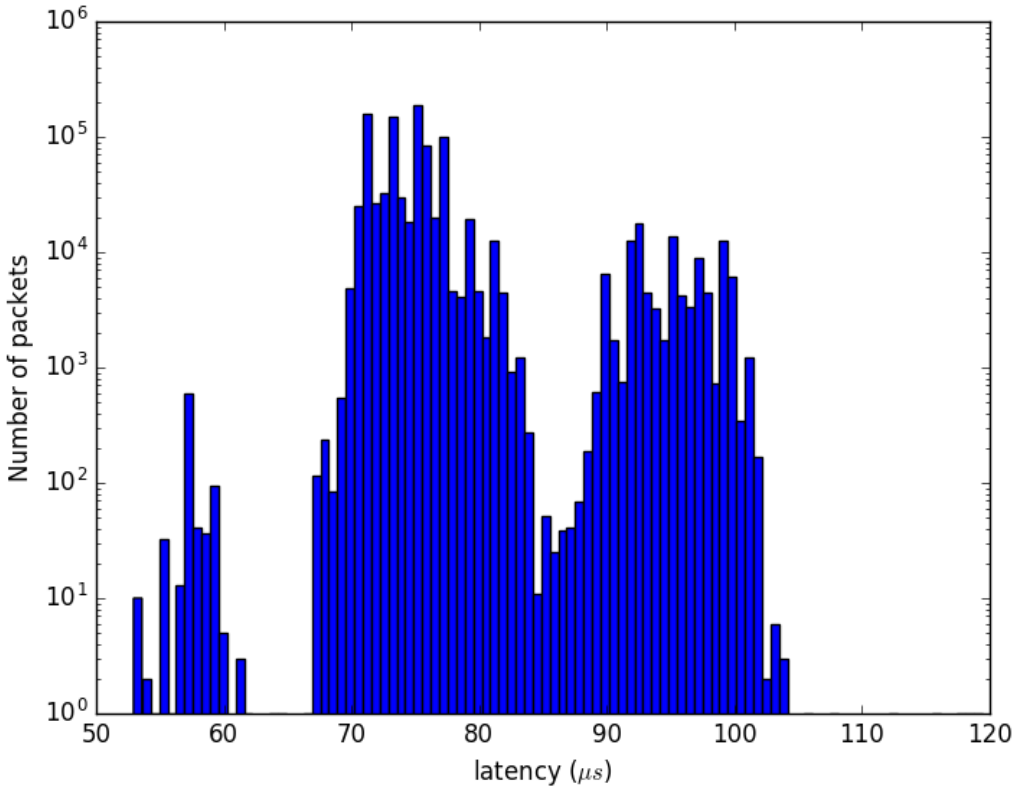
MPI: latency

Mean latency vs. message size



MPI: jitter

Jitter, 1kB messages



Pipeline middleware downselect matrix

Criterion	Weighting	Technology	Remarks	Score	Weighted score
DS-MW-1 Reliability	3	ZeroMQ	No guaranteed delivery	0	0
		MPI	Reliable QoS available	3	9
DS-MW-2 Latency	3	ZeroMQ	Unable to meet requirement	0	0
		MPI	Required performance achieved in testing	3	9
DS-MW-3 Jitter	3	ZeroMQ	Unable to meet requirement	0	0
		MPI	Required performance achieved in testing	3	9
DS-MW-4 Throughput	3	ZeroMQ	Required performance achieved in testing	3	9
		MPI	Required performance achieved in testing	3	9
DS-G-1 Cost	1	ZeroMQ	Available free/open source	3	3
		MPI	Available free/open source	3	3
DS-G-2 Ease-of-use	1	ZeroMQ	Commensurate with facilities provided	2	2
		MPI	Commensurate with facilities provided	2	2
DS-G-3 Long-term support	2	ZeroMQ	Single supplier; commercial support available	2	4
		MPI	Several implementations available, and very widely used.	2	4
DS-G-4 Standards compliance	2	ZeroMQ	No standard	0	0
		MPI	De-facto HPC standard	1	2
DS-G-5 Familiarity	2	ZeroMQ	Expertise in consortium	1	2
		MPI	Expertise in responsible partner	2	4
DS-G-8 Source of supply	2	ZeroMQ	Single supplier	1	2
		MPI	Multiple implementations	3	6
Overall Score		ZeroMQ			22
		MPI			57

Conclusions (1)

- ZeroMQ: unsuitable for real-time pipeline
  - Excessive latency: ~ 50% total budget, probably owing to internal buffering and message aggregation
- MPI: latency and jitter adequate
  - ~ 5% of latency budget, for small messages
  - Hence, limited number of network hops allowed
  - Hence, some constraints on implementations using MPI
    - number of hops, message size

Criteria: telemetry middleware

ID	Criterion	Description	Weighting
DS-MW-5	Publish/ subscribe paradigm	Middleware must abstract the distribution of data from producers to consumers via the publish/subscribe pattern, wherein the middleware rather than user code is responsible ensuring all published are sent to all consumers which have subscribed to said data..	3
DS-MW-6	Single writer/multiple reader paradigm	Middleware must support distribution of data to multiple consumers via a single publication action by the producer.	3
DS-MW-7	Service discovery	Middleware should provide facilities for producers and consumers to locate one another rather than requiring the user to do so programatically.	2
DS-MW-8	Location transparency	Related to DS-MW-7, middleware should abstract the location (IP address,etc) of services from user code, allowing these to be addressed by name rather than location.	2
DS-MW-9	Filtering	Middleware should allow subscribers to apply filtering rules to a data stream to select from the published data (eg. time-based, key/value based, frequency, etc	2

Telemetry middleware downselect matrix

Criterion	Weighting	Technology	Remarks	Score	Weighted score
DS-MW-1 Reliability	3	ZeroMQ	No reliable delivery	0	0
		DDS	Reliable QoS available	3	9
DS-MW-4 Throughput	3	ZeroMQ	Cannot satisfy most demanding case	1	3
		DDS	Cannot satisfy most demanding case	1	3
DS-MW-5 Publish/ subscribe	3	ZeroMQ	Supported	3	9
		DDS	Supported	3	9
DS-MW-6 Single writer/multiple reader	3	ZeroMQ	Supported	3	9
		DDS	Supported	3	9
DS-MW-7 Service discovery	2	ZeroMQ	Broker exists as separate project, not part of ZeroMQ	0	0
		DDS	Broker, or broker-less discovery service	3	6
DS-MW-8 Location transparency	2	ZeroMQ	Address of peers must be known	0	0
		DDS	Entirely transparent broker-less operation possible	3	6
DS-MW-9 Filtering	2	ZeroMQ	Not supported by middleware	0	0
		DDS	Wide range of filtering options	3	6
DS-G-1 Cost	1	ZeroMQ	Free/open source	3	3
		DDS	Free/open source and free commercial licences available	3	3
DS-G-2 Ease-of-use	1	ZeroMQ	Commensurate with facilities provided	2	2
		DDS	Commensurate with facilities provided	2	2
DS-G-3 Long-term support	2	ZeroMQ	Single supplier, commercial support available	1	2
		DDS	GPL and several commercial suppliers; commercial support available	3	6
DS-G-4 Standards compliance	2	ZeroMQ	Proprietary, no standard	1	2
		DDS	OMG standard	3	6
DS-G-5 Familiarity	2	ZeroMQ	Expertise in consortium	2	4
		DDS	Considerable expertise in responsible partner	3	6
DS-G-8 Source of supply	2	ZeroMQ	Single supplier	1	2
		DDS	Multiple suppliers	3	6
Overall Score		ZeroMQ			36
		DDS			77

Telemetry: derived throughput requirements

Throughput requirements for undecimated telemetry data types:

Source	Remarks	Frame size (bytes)	Throughput (MB/s)
Raw pixels	800x800 pixels @ 800Hz, 2 bytes/pixel	1,280,000	1024
Calibrated pixels	800x800 pixels @ 800Hz, 4 bytes/pixel	2,560,000	2048
Slopes	74x74 subapertures @800Hz, 2x4 bytes/subap	43,808	35.05
Mirror demands	75x75 actuators, @800Hz, 4 bytes/actuator	22,500	18

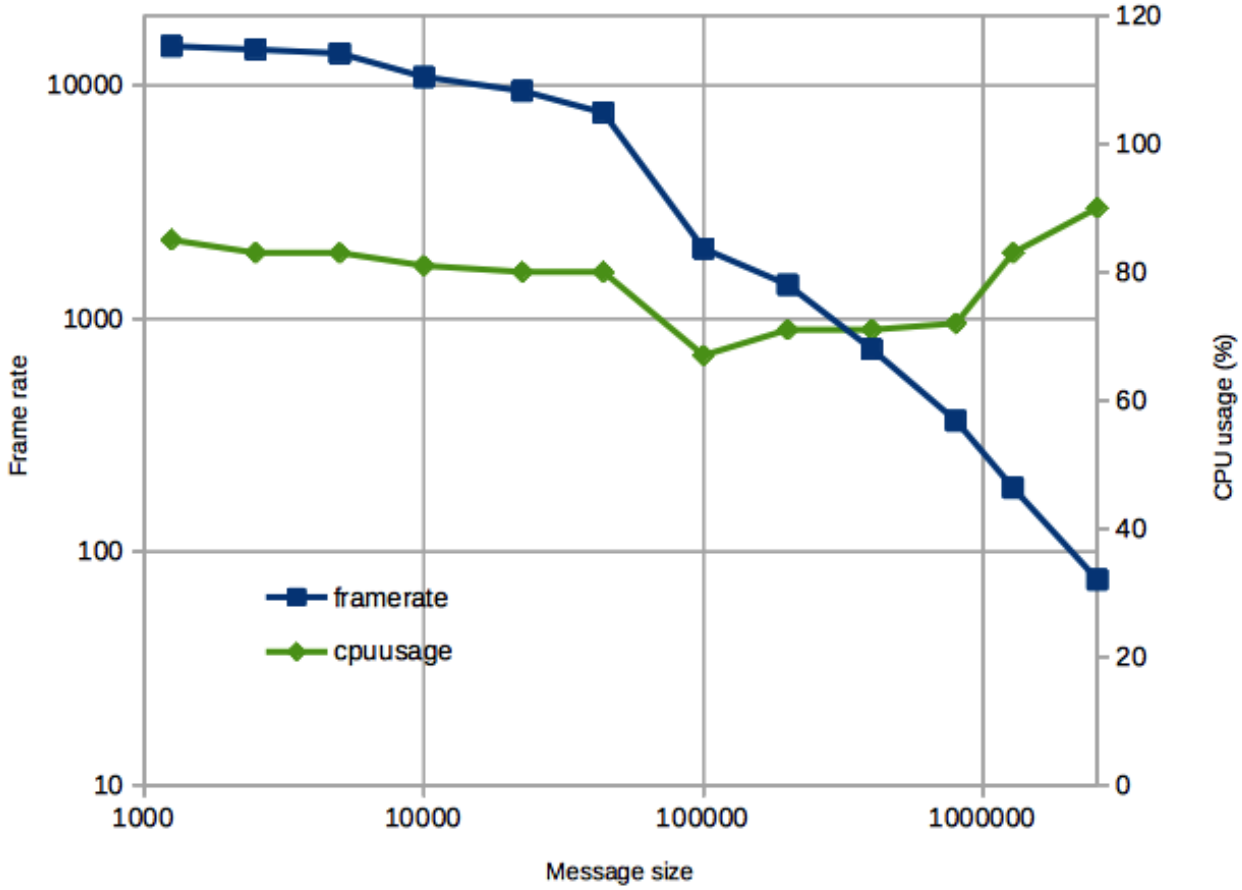
DDS test environment

- RTI DDS
  - RTI provided throughput test software
  - Sends data as fast as possible
  - Various representative message sizes
- Intel Phi with 16GB HBM
- 40Gbe connection, source → switch → consumer
  - Also tested shared memory transport, single node

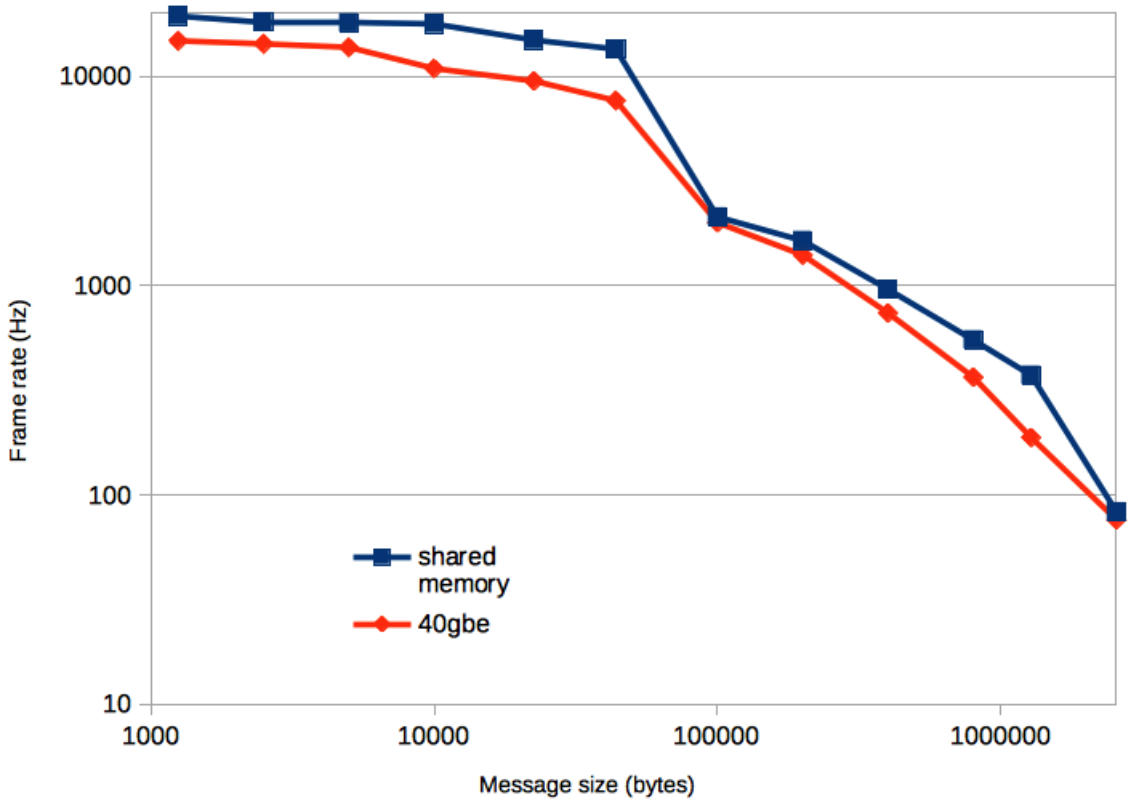
DDS:measured throughput

Message size (bytes)	40 Gbe			Shared memory		
	Frame rate (Hz)	Thruput (MB/s)	CPU usage (%)	Frame rate (Hz)	Thruput (MB/s)	CPU usage (%)
22500 (mirror demands)	9500	214	80	14,900	335	101
43808 (slopes)	7660	336	68	13,500	591	100
1,280,000 (raw pixels)	188	241	83	373	477	71
2,560,000 (calib. pixels)	76	195	90	83	212	37

DDS max frame rate, CPU usage, Intel Phi, 40Gbe



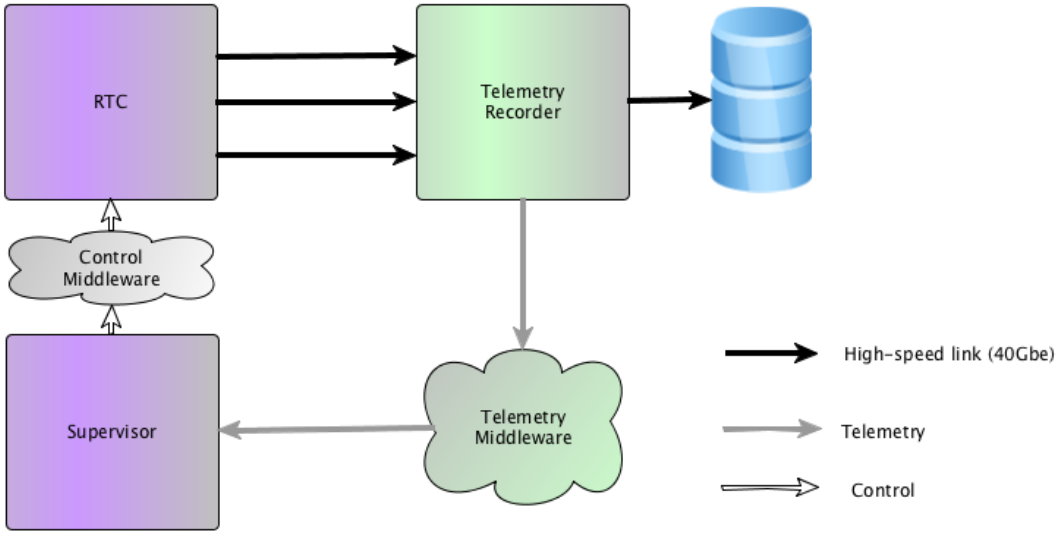
DDS max framerate, 40Gbe vs. shared memory (Xeon Phi)



Conclusions (2)

- DDS provides the required facilities
- Inadequate throughput performance for undecimated pixel telemetry
- Undecimated slopes/DM demands are supportable
- CPU-limited for smaller messages: shared-memory test achieved higher frame rates, saturated CPU core

Telemetry Solution



- GF requirement: capture 10s of contiguous real-time telemetry
- High-speed links from RTC to telemetry recorder
- Telemetry recorder publishes to DDS, decimated as needed
- Supervisor subscribes via DDS
- Supervisor uploads to RTC using control middleware

Criteria: Control Middleware

ID	Criterion	Description	Weighting
DS-MW-1	Reliability	The middleware should be able to guarantee delivery of uncorrupted data, or at the least, detect and signal non-delivery or data corruption.	3
DS-MW-7	Service discovery	Middleware should provide facilities for producers and consumers to locate one another rather than requiring the user to do so programatically.	2
DS-MW-8	Location transparency	Related to DS-MW-7, middleware should abstract the location (IP address,etc) of services from user code, allowing these to be addressed by name rather than location.	2
DS-MW-10	Request/reply paradigm	Control middleware should support a request/reply (aka. command/response) synchronous interaction pattern, RPC-like or similar in nature.	3

Control middleware downselect matrix

Criterion	Weighting	Technology	Remarks	Score	Weighted score
DS-MW-1 Reliability	3	ICE	Limited reliability	1	3
		DDS	Reliable QoS available	3	9
DS-MW-7 Service discovery	2	ICE	Via object broker	1	2
		DDS	Broker or broker-less discovery service	3	9
DS-MW-8 Location transparency	2	ICE	Proxies provided by broker	2	4
		DDS	Entirely transparent broker-less operation possible	3	6
DS-MW-10 Request/reply	3	ICE	Supported	3	9
		DDS	Supported	3	9
DS-G-1 Cost	1	ICE	Free/open source	3	3
		DDS	Free/open source and free commercial licences available	3	3
DS-G-2 Ease-of-use	1	ICE	Commensurate with facilities provisded	2	2
		DDS	Commensurate with facilities provided	2	2
DS-G-3 Long-term support	2	ICE	Single supplier. Commercial support available	1	2
		DDS	GPL- and several commercial suppliers. Commercial support available	3	6
DS-G-4 Standards compliance	2	ICE	Proprietary	0	0
		DDS	OMG standard	3	6
DS-G-5 Familiarity	2	ICE	No expertise in consortium	1	2
		DDS	Considerable expertise in responsible partner	3	6
DS-G-8 Source of supply	2	ICE	Single supplier	1	2
		DDS	Multiple suppliers	3	6
Overall Score		ICE			29
		DDS			62

Conclusions

Middleware domain	Technology	Weighted score
Low latency	MPI	57
	ZeroMQ	22
Telemetry	DDS	77
	ZeroMQ	36
Control	DDS	62
	ICE	29