

End-to-End Diagnostics

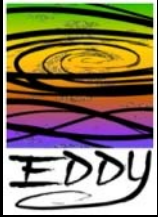
Progress to date and diagnostic use case efforts in the areas of Email, networking, environmental management and security.

Internet2/Fall Member Meeting 12/06

Chas DiFatta (chas@cmu.edu)

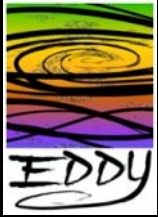
Mark Poepping (poemming@cmu.edu)

Walter Wong (wcw@cmu.edu)



Outline

- Present diagnostic areas of focus
 - Email
 - Environmental
 - Security
 - Network
- Current Activities



Scope Bounding

Machine Information

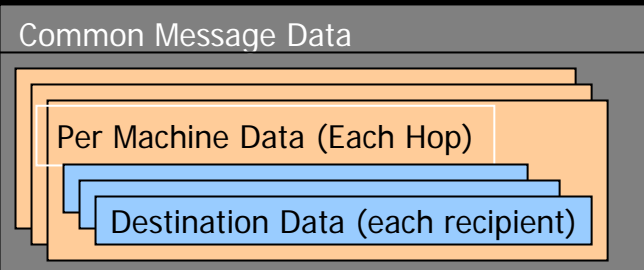
“Standard” Machine Information

Run Queue load; Summary I/O, etc

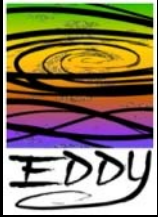
“Email Server” Information

Run Queue load; Summary I/O, etc

Message Information

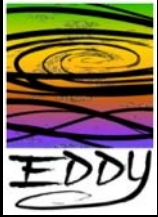


- This discussion is about the right side of the white line: Information about the message
- Data on the left side of the green line will be considered later. Information about the machines processing the message.



Data Source

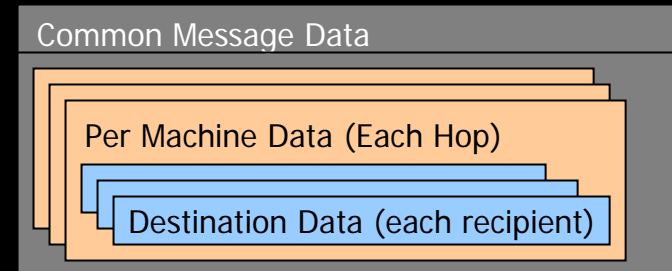
- Raw data comes from the syslog output of sendmail
 - Could be generalized to other transport tools (e.g. postfix)
- Sendmail log entries generated after specific actions occur
 - This logging created to see what *sendmail is doing* not what *sendmail is doing with the message*
- Messages usually traverse multiple machines
- To get information the diagnostician will generally need to
 - do multiple grep passes through large files (one to find the sendmail queue id and then using the queue-ids to find the other lines)
 - connect to multiple machines to do the search (unless there is central syslogging – which has its own issues)

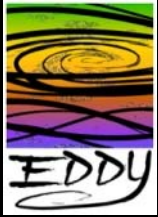


Data Model for Email

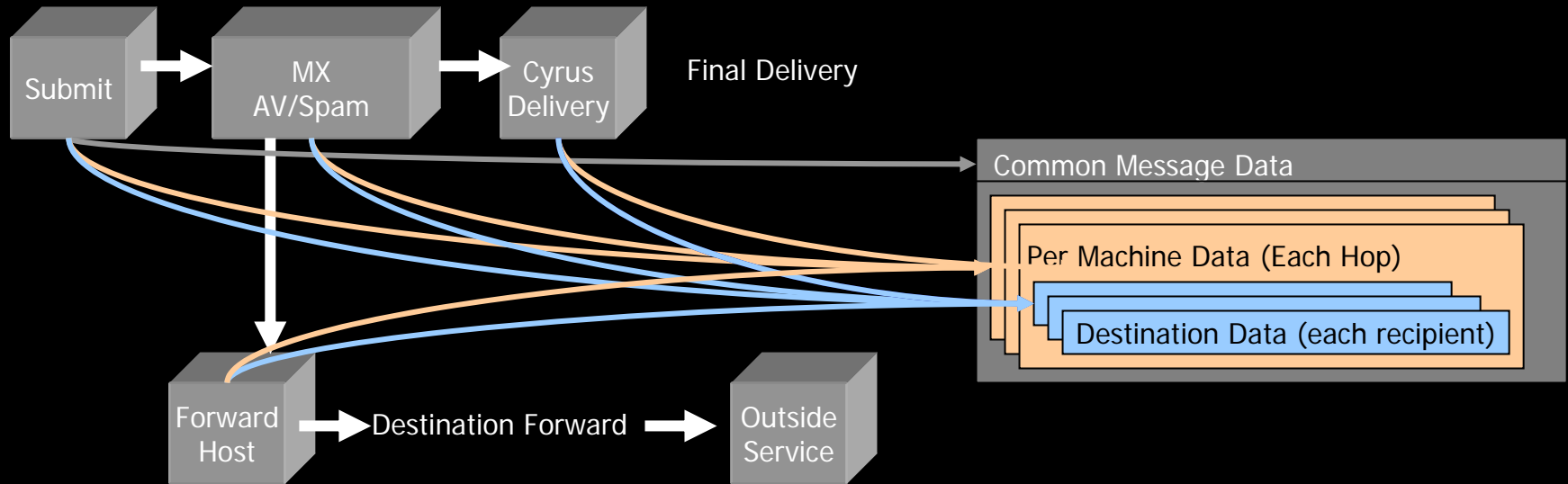
To focus on the message, we create a *message flow record* that contains:

- Data that is common for the message across transit hosts (e.g. from / message id)
- Capture data for each hop the message took – per machine data (e.g. local QueueID)
- Within each machine record, capture information about each destination (e.g. specific destination, message sent? tmp fail?)

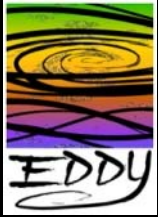




Message Progression

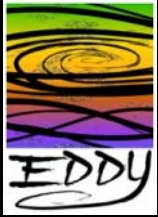


- Pulling apart that last diagram, we see how the flow is built through progression of the message
- This example represents a single message, submitted locally, with two destinations, one local and one which will be forwarded off-site, outside messages hit the MX first, of course.
- At the first hop, we see the message for the first time, create the message record with the common data and add the per-machine data for the first hop (e.g. destinations)
- At the second hop (the MX host), we add the per-machine data plus the anti-spam and AV info
- We then see it split, the top path results in final delivery
- For the bottom path, the destination isn't actually the final stop (e.g. .forward or sieve), we see that it was forwarded off-site, but have no further information (outside service)



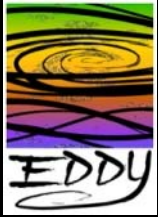
Common Message Data

- Primary cross machine correlation keys
- Envelope From: will be constant message-id may not exist and so would be added by first MTA
- Possible other items (data not available by default)
 - Size of the body / (hash of the body)
 - Subject (hash of subject)
 - Body date
 - Body to/from/subject



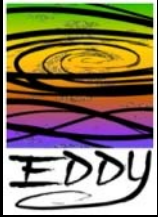
Per Machine Data

- Typical focus of mail log processing tools
- Data Components
 - sendmail data
 - Queue-id; received timestamp; relay host (source host); auth information
 - milter/support data
 - spam check; AV; grey list foo
- Per recipient component
 - number recipients
 - *<number recipients>* Recipient Information Records



Destination Data

- At least one per machine information record
- Typically *to=* line in sendmail log
- Time to deliver; delivery location; delivery status



Simple Message Flow Example

Raw Log (edited for clarity)

```
Feb 21 07:06:02 [18350]: from=<richard@yahoo.com>, size=27751, class=0, nrcpts=1,  
msgid=<000001c62a4e$7b3c4180$0100007f@neska>, proto=SMTP, daemon=MTA, relay=c-24-18-9-  
72.hsd1.wa.comcast.net [24.18.9.72]  
Feb 21 07:06:02 [18350]: Milter add header: X-Spam-Warning: 99% (... more spam info ...)  
Feb 21 07:06:02 [18350]: greylst: addr 24.18.9.72 from <richard@yahoo.com> rcpt <wcw@andrew.cmu.edu>:  
autowhitelisted for 24:00:00  
Feb 21 07:06:02 [18350]: Milter add header: by milter-greylst-2.0 [...]  
Feb 21 07:06:02 [18363]: to=<wcw@andrew.cmu.edu>, delay=00:00:01, xdelay=00:00:00, mailer=cyrusmurder,  
pri=147751, relay=local host, dsn=2.0.0, stat=Sent
```

Common Message Data

From: richard@yahoo.com

Message-Id: 000001c62a4e\$7b3c4180\$0100007f@neska

Per Machine Data

machine: mx8.andrew.cmu.edu

queue-id: k1LC60xK018350

relay host: relay=c-24-18-9-72.hsd1.wa.comcast.net
[24.18.9.72]

Spam tag: X-Spam-Warning: 99% [...]

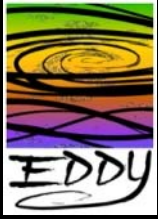
GreyList: X-Greylst: Delayed for 00:00:02 [...]

Destination Data

To: wcw@andrew.cmu.edu

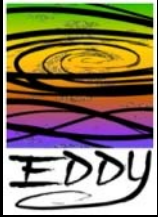
Delay: 00:00:01 / xdelay: 00:00:00

Status: Sent



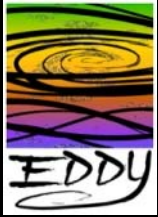
Ongoing Issues

- Message progression through mailing list software
 - May change our primary keys (message-id / from)
- Duplicate message-id
- Integration / correlation with CERs
 - Intermediate Event model



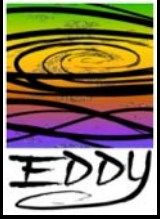
Recap on Focus

- The focus is on the message
 - Tell me all the messages sent to help-desk in this time range since the help-desk software screwed up and dropped all incoming on the floor.
 - User X says they aren't receiving email? Show me the email to user X.
 - User X says user Y isn't receiving email. What can I find out?
- The focus is not on the server (yet). Can't answer:
 - why did messages on server A take longer to deliver than message Y



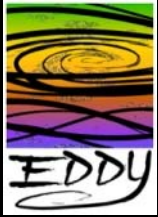
Outline

- Present diagnostic areas of focus
 - Email
 - Environmental
 - Security
 - Networking
- Current Activities

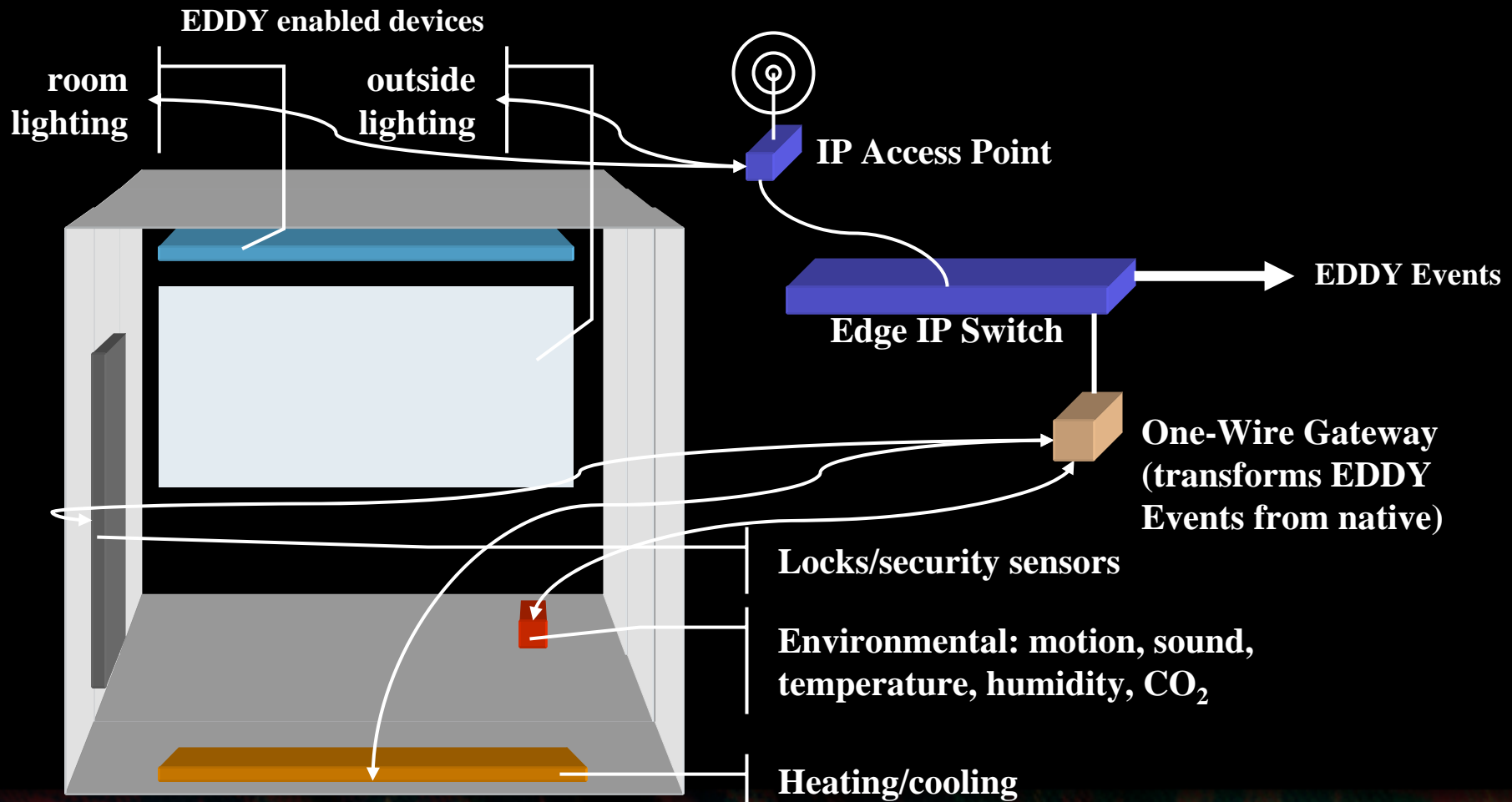


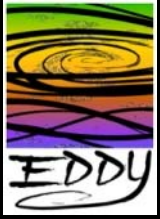
Questions Answered

- When was the temperature above 78 degrees in Hamburg Hall on Nov 16th?
- Approximate the occupancy of the 7th floor between 2 and 4pm?
- Who was in the office last night when the motion alarm when off and what was the temperature 15 min after?



Room Events



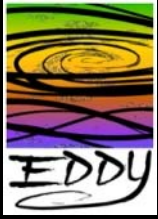


Challenges and Opportunity

Challenge: Embedded systems have limited resources

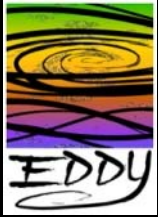
- Memory
- Development platforms (limited feature set of porting language)

Opportunity: developers want a simple service mechanism to create CER's

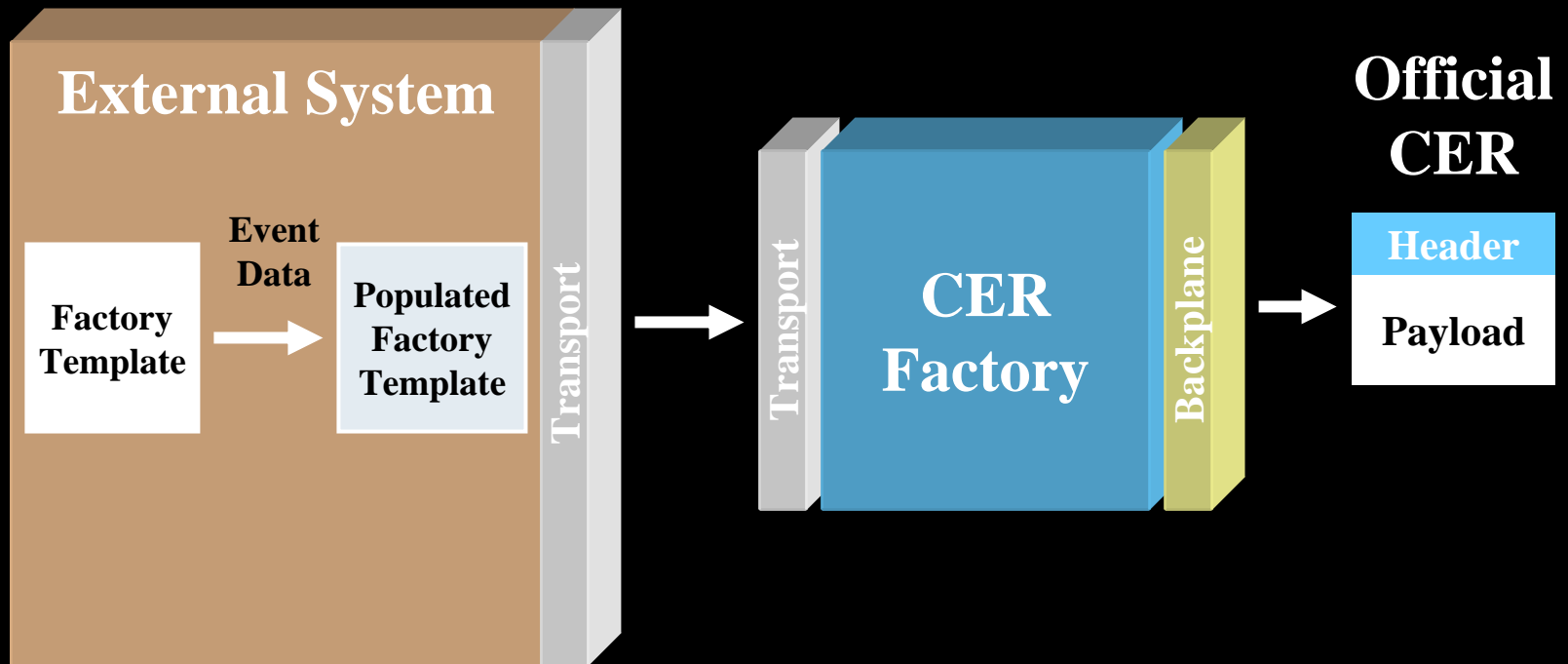


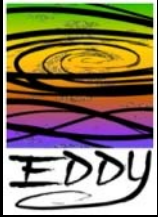
CER Factory Concept

- Facility to easily create CERs from a limited set of event attributes
- Simplify using the backplane for non-Java developers
- Reduce the barrier to adoption for embedded system developers
- Provide a wide array of transport options (SSL, HTTP, SOAP, native sockets, etc.)



CER Factory Architecture

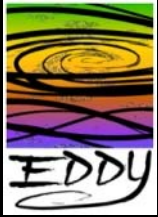




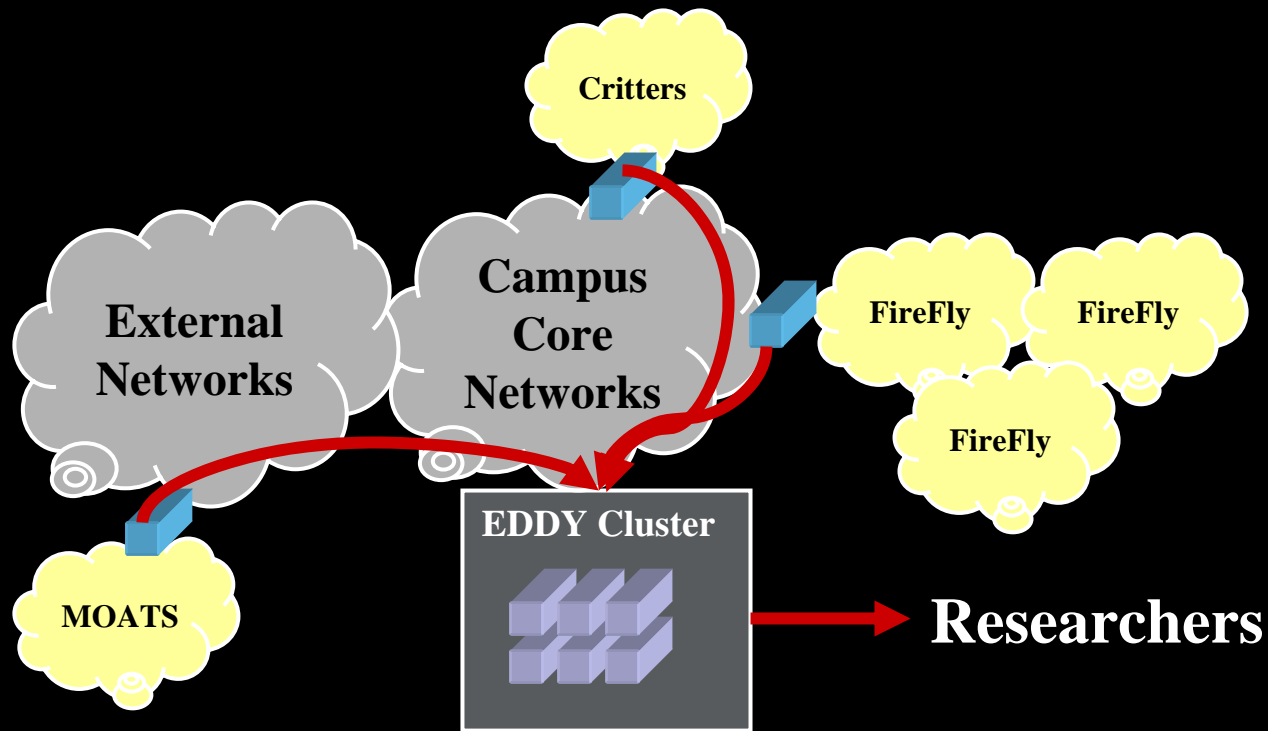
CER Factory Template Example

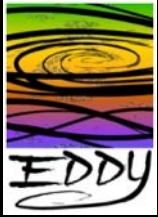
Temperature sensor event as a Factory Template:

```
<?xml version="1.0" encoding="UTF-8"?>
<cerFactory>
  <eventInfo.oid>1005</eventInfo.oid> <!-- CBPD -->
  <eventInfo.eventHostname>sensor1.net.cmu.edu</eventInfo.eventHostname>
  <eventInfo.eventClass>5</eventInfo.eventClass> <!-- Environmental event -->
  <eventInfo.warningLevel>7</eventInfo.warningLevel> <!-- Informational warning level -->
  <dataPayload.payloadType>2</dataPayload.payloadType> <!-- Cooked record -->
  <dataPayload.payload>
    <cbpd>
      <cbpd-1.0.0>
        <temperatureSensor>
          <temperatureSensorSetPoint>74.61</temperatureSensorSetPoint>
        </temperatureSensor>
      </cbpd-1.0.0>
    </cbpd>
  </dataPayload.payload>
</cerFactory>
```



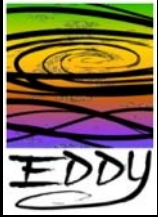
Campus Architecture





Outline

- Present diagnostic areas of focus
 - Email
 - Environmental
 - Security
 - Networking
- Current Activities

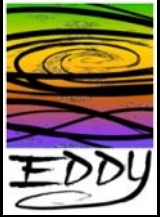


Security Efforts

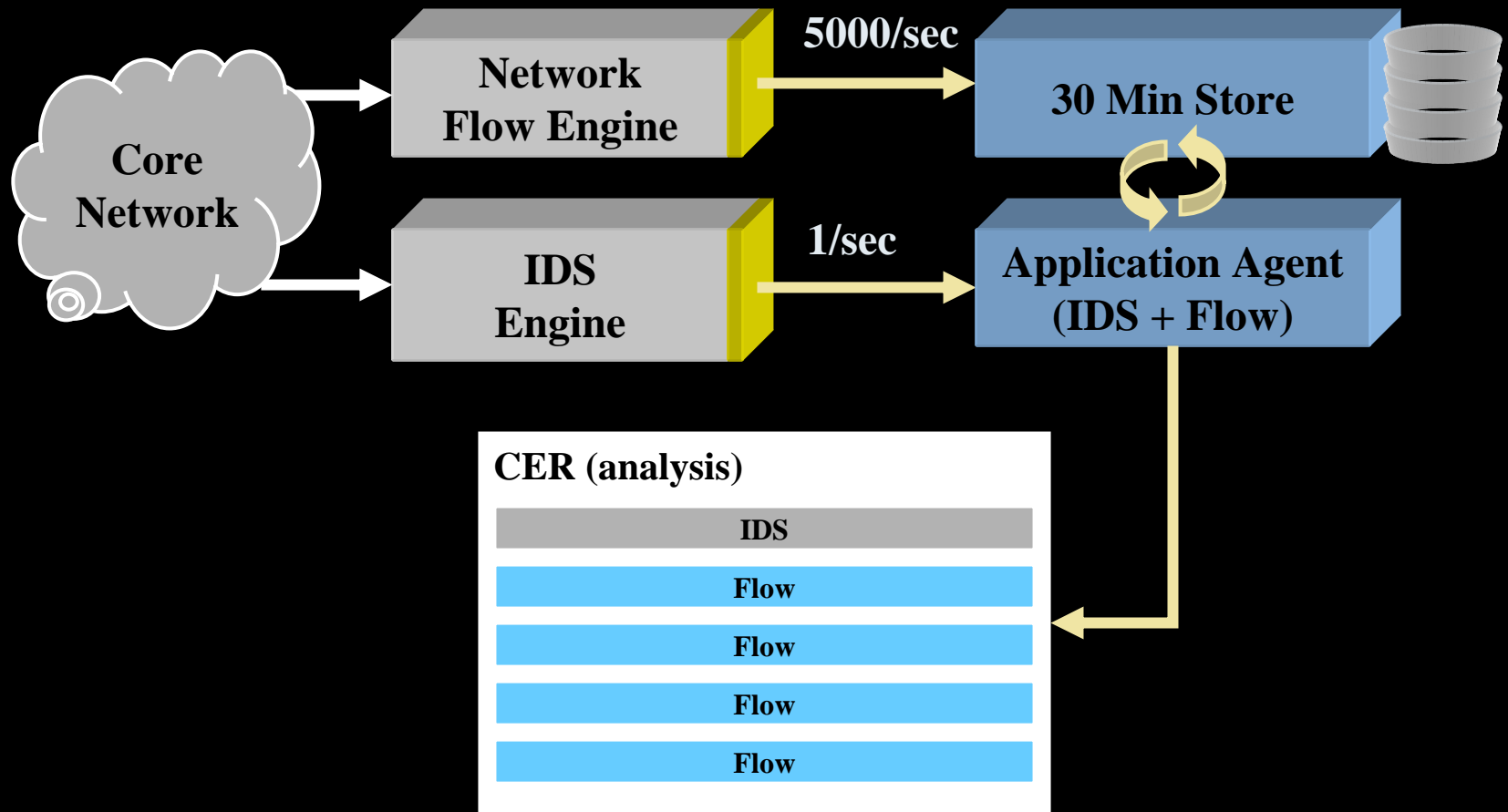
Problem: IDS events need to be verified using other types of events (network, system, application)

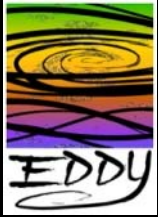
Example: If an attack on the SMTP servers was detected by the IDS, did the server logs reflect anything interesting?

Solution: Combine IDS with network flow events.



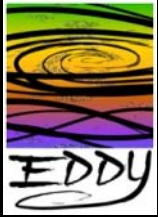
IDS + Flow Events





Outline

- Present diagnostic areas of focus
 - Email
 - Environmental
 - Security
 - Networking
- Current Activities



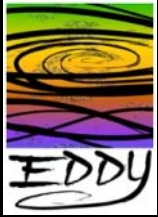
Network Efforts

Managing Scale and Privacy

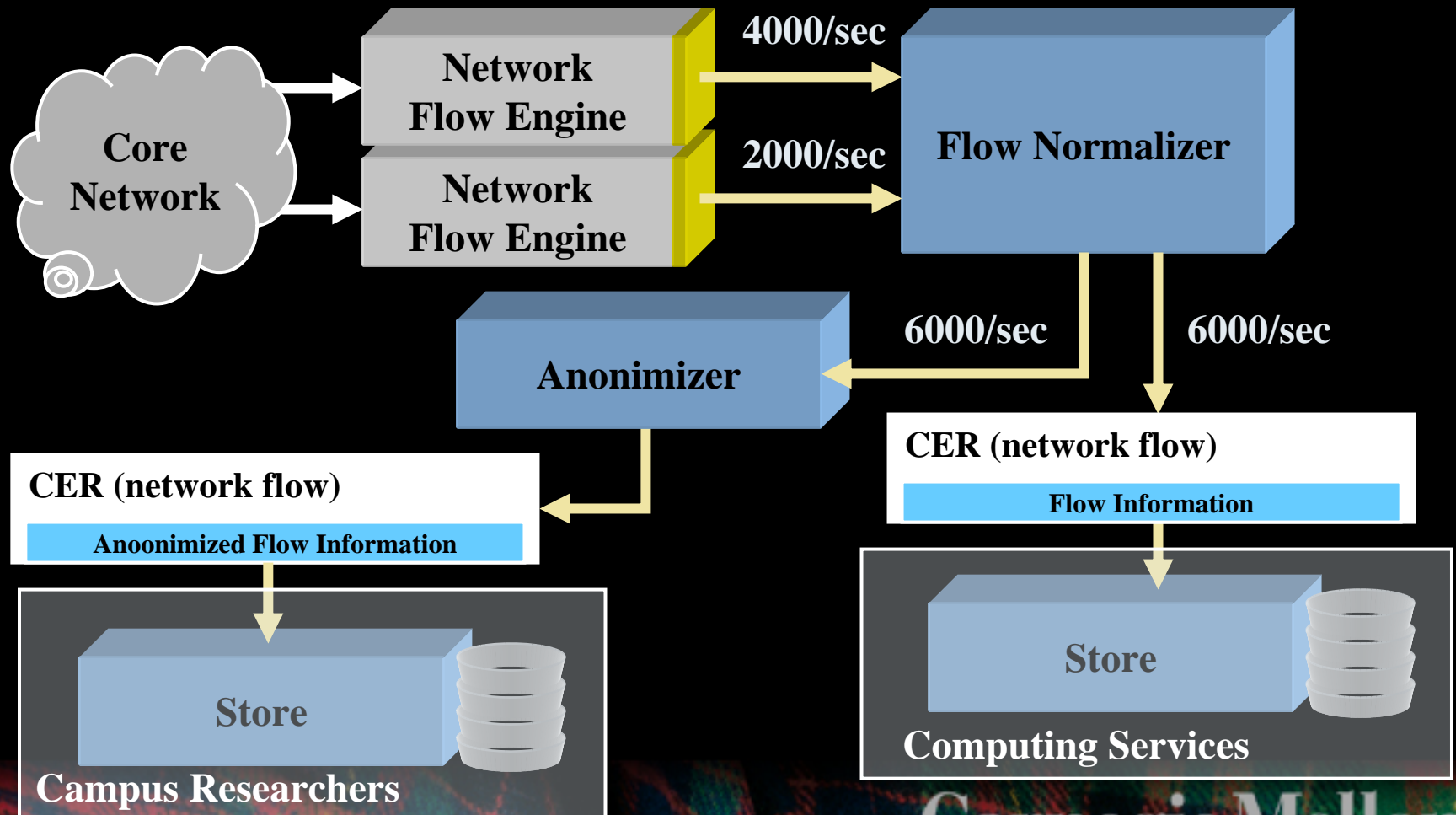
Problem: network flow events used for management are highly desired by the research faculty but contain sensitive data.

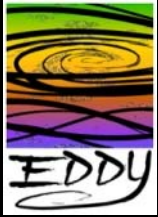
Solution:

- Strip out only data that is needed for management and route to network management group
- Anonymize source and destination addresses and strip out any data that is not permitted by the IRB.



1st Order Flow Events





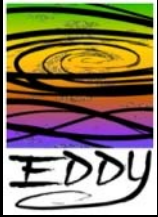
Network Efforts

Multiple Views Into Data

Problem: Need to get a real-time view on egress network usage from different perspectives.

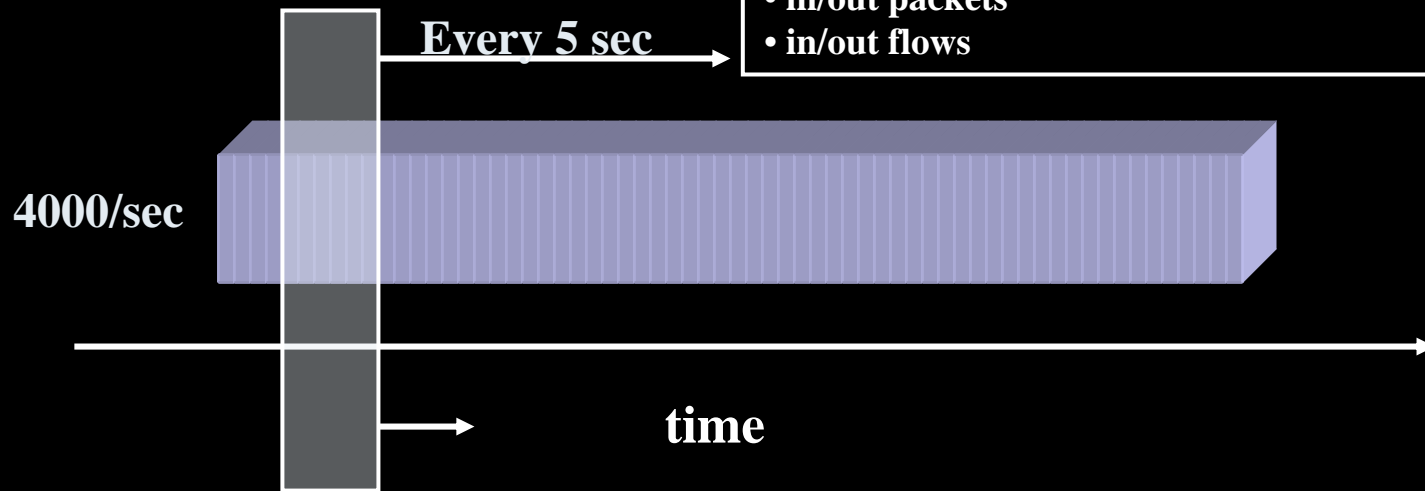
Example: Why is the Internet2 link overloaded?
What is accounting for most of the traffic and what is it's usage?

Solution: Aggregate flow events from different perspectives (flows, bytes, packets)



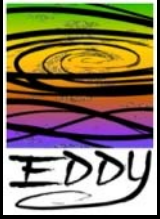
2nd Order Flow Events

One minute sliding window
of top 100 talker information

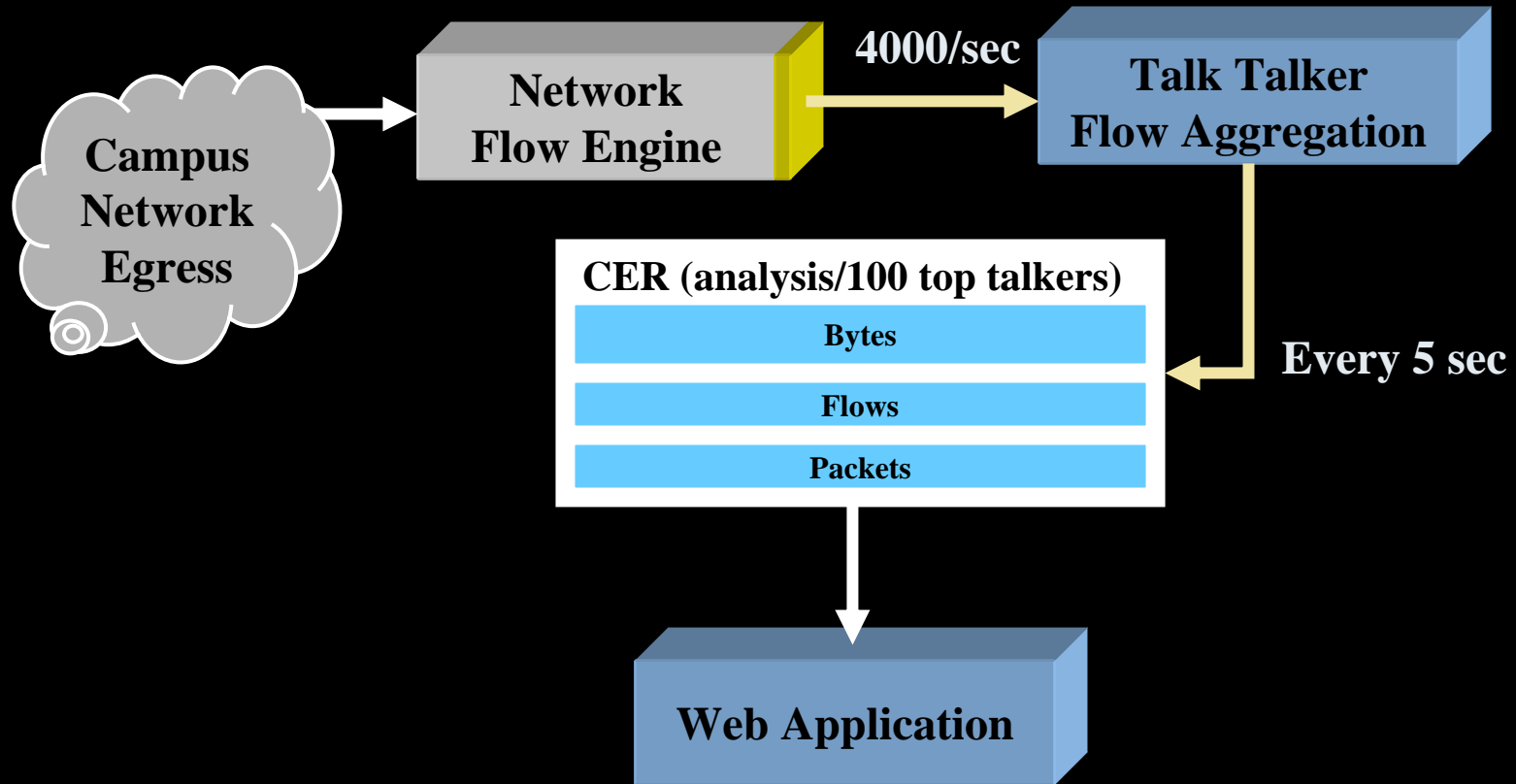


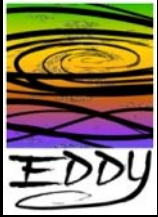
For each top talker (100 x 3 [bytes, flows, packets])

- IP Address
- Service ports
- in/out bytes
- in/out packets
- in/out flows



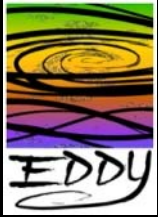
2nd Order Flow Events





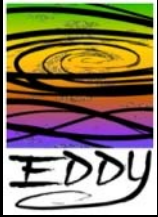
Outline

- Present diagnostic areas of focus
 - Email
 - Environmental
 - Security
- Current Activities



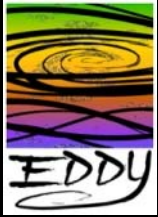
Future Development Directions

- Simplify the injection of new events
 - Add new types of CER
 - Toolkits for other languages (e.g. Perl)
- Mature the CER- Version 2
 - New Features: dynamic, performance focused
 - Standardize: Working with IBM on format and transport
- Develop the query and control channel
- Add new high value agents (e.g. storage, discovery)



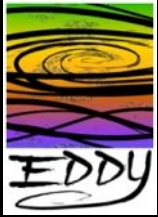
Effort Status

- Development
 - Initial release (Munster 0.5) targeted at developers - 4/1/06
 - EDDY Agent Framework, Sample EDDY Agents, Agent Manager
 - Supplemental release (Sushi 0.5.1) targeted at network managers -8/16/06
 - Framework enhancements, Normalizers, TopNetworkTalker Application
 - Supplemental release (Murphy 0.5.2) beta released 10/12/06
 - Additional framework enhancements, Normalizers, internal diagnostics, CER factory
 - Supplemental release (0.6) targeted at Email and security diagnosticians
 - More Normalizers, Email and security applications
- Outreach
 - Involving others in the development process
 - Expand to other use cases external to CMU
 - Working with industry leaders on proposed standards and methods
- Support
 - Initially sponsored by the National Science Foundation under the NSF
 - CyDAT (Cyber-center for Diagnostics Analytics and Telemetry)
 - Soliciting partners in industry



Want to Learn More?

www.cmu.edu/eddy



End-to-End Diagnostics

Progress to date and diagnostic use case efforts in the areas of Email, networking, environmental management and security.

Internet2/Fall Member Meeting 12/06

Chas DiFatta (chas@cmu.edu)

Mark Poepping (poemming@cmu.edu)

Walter Wong (wcw@cmu.edu)