# Poster Abstract: Learning from Sensor Network Data

Matthias Keller, Jan Beutel, Andreas Meier, Roman Lim, and Lothar Thiele
Computer Engineering and Networks Lab, ETH Zurich
8092 Zurich, Switzerland
matthias.keller@tik.ee.ethz.ch

## Abstract

Within the PermaSense project, two wireless sensor networks have been deployed for a long-term operation in the Swiss Alps. For enabling state-of-the-art permafrost research based on the collected data, highest possible data quality and yield have to be ensured. But, the operation of wireless sensors networks remains a hard research problem. Firstly, deployed wireless sensors networks are subject to continuous changes. Second, there are scenarios that can only be tested in the field as the capabilities of testbeds are too limited. Basically, it is not possible to test for many months before deploying in the field. In this poster, we present an analysis of our data that has been collected over nine months. In addition to describing our system design and methods, we also share our experiences from discovered severe incidences.

## Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design; B.8.2 [**Performance and Reliability**]: Performance Analysis and Design Aids

## General Terms

Performance, Design, Reliability, Verification

## Keywords

Wireless Sensor Networks, Environmental Monitoring, Long Term, Data Analysis

## 1 Introduction

The PermaSense project [1] strives for collecting geophysical data with wireless sensor networks. There are currently two deployments that are exposed to the harsh environmental conditions of high-altitude areas for more than 12 months to this day.

Data quality and yield is a serious concern since the emergence of wireless sensor networks. Starting with the early work of Szewczyk et al. [2], there is also evidence of extensive investigations in other deployments such as the work of Werner-Allen et al. [3].

During our planned operation of at least three years, our goal is to ensure the highest possible data quality and yield. In this poster, we describe an analysis of real deployment data and the results gathered from this analysis so far.


*SenSys'09*, November 4–6, 2009, Berkeley, CA, USA.
ACM 978-1-60558-748-6

## 2 System Design

The PermaSense deployment is built on TinyNode sensor nodes running TinyOS-based PermaDozer [4]. Data from the sensor nodes is collected at the base station consisting of an access node that is connected to an embedded PC. From the base station, IP networks (WLAN, GSM) are used for transmitting the collected data to the backend server running Global Sensor Network.

All sensor nodes excluding the access node generate a set of five packets every two minutes. Each packet contains a sequence number and a timestamp. The sequence number is local to a particular sensor node and incremented with each packet. The most recently used sequence number is non-volatile in case of resets when there is no power loss.

One of the five packets is used for health information. This information consists of the node uptime, battery voltage, temperature and humidity within the enclosure plus information from the routing protocol about the network tree topology. More precisely, the address of the parent node, the number of hops towards the sink, the number of children and the number of pending messages in the outgoing queue are sent.

For maximizing the data yield, two data recovery provisions are available. Firstly, sensor nodes are equipped with SD memory cards that are used for storing generated packets. Second, packets are logged at the base station when passing the interface between the access node and the embedded PC.

## 3 Data Analysis

The data analyzed from the Matterhorn deployment [4] was generated between September 2008 and May 2009. Within this time span, over 14 million packets in total have been generated at 13 sensor node positions plus two positions for relay nodes.

*Collected data is not always immediately usable.* Due to the ongoing development of software, hardware and models during operation, we have to deal with different data storage locations and data formats.

Our extensive cleaning process has to tackle three main problems. First, all data must be converted to the most recent data format. Second, removal of multiple entries of the same packets needs to be carried out. These duplicates are related to former software bugs and non-deterministic system behavior due to partial node failures. This step of the process is directly connected of the most elaborate third step

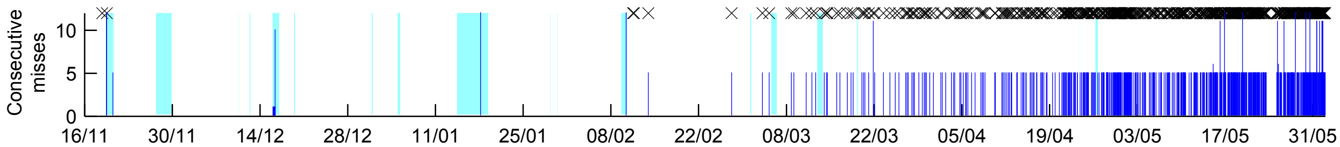


**Figure 1. Performance plot showing the lost packets (bars) and the highly correlated node resets (crosses) over a period of 28 weeks starting from November 2008. The shaded areas mark time frames in which infrastructure problems at the base station stopped data collection.**

in which the data is ordered. As the system does not guarantee that packets arrive in the same temporal order as the packets have been generated, packet sequence numbers and timestamps are used for both removing duplicates and reconstructing the proper order.

Not only learning that data cleaning can be really extensive, it is also noticeable that meta information for tracking all kinds of changes is really important.

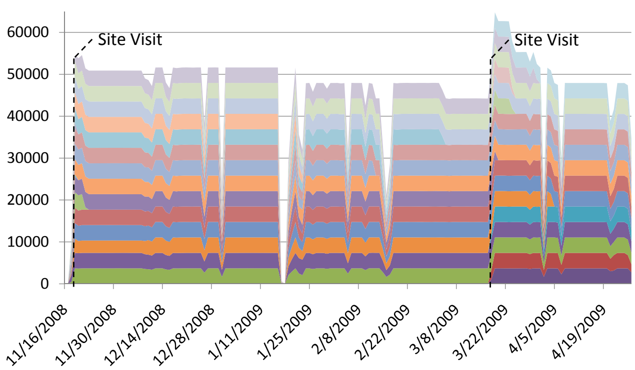


**Figure 2. Number of received packets per day from September 2008 to April 2009. Each data series corresponds to a sensor network node.**

*Data collection does not stop at the access node and includes safe storage at the backend.* Figure 2 shows the number of packets that have been received per day. It only includes directly received packets without recovery from the described backup facilities. While the number of nodes decremented due to node deaths over time, basically no data was lost in the sensor network in the time from November 2008 to March 2009. While 95 percent of all generated packets were directly received from 12 of up to 15 nodes over the whole time span, missing packets from healthy nodes relate to outages of the backend server and the base station, but not to the sensor network itself. For instance, data from an outage in mid of January can be restored from recovered memory cards, data corresponding to the gap in mid of February is available on both described backup facilities. In the future, a newly implemented layer for safe, acknowledged transmission of data from the base station to the backend will help to increase the portion of directly received data.

*Long-term effects can only be observed in the field.* Around 8.400 packet losses occurred between July 2008 and June 2009. It turns out, that exactly five consecutive packets were lost in over 70 percent of all cases. Further investigation shows, that a loss of five consecutive packets corresponds to a node reset that occurred between data acquisition and storage with transmission phases.

Figure 1 shows the temporal distribution of node resets for a representative node. Apparently, there is a performance degradation starting in mid of March 2009. While Figure 1 shows only one node, this effect applies to all nodes being deployed in September 2008. In turn, the two nodes that have been replaced in March 2009 are not affected.

Up to 40 resets per node and day can be observed in the degradation phase. After physically recovering some of the affected nodes, it turned out that these resets are caused by software due to a task queue overrun. This happened due to a performance leak in the data storage implementation. Concretely, the execution time of the task used for reading and writing data from the SD memory card increased with the amount of stored data. It becomes apparent, that this kind of long-term error can hardly be observed while testing on a testbed for a couple of weeks.

While the current progress of the data analysis was really insightful concerning the detected performance, we are still interested in a more detailed analysis of the network performance.

## 4 Acknowledgments

The work presented was supported by NCCR-MICS, a center supported by the Swiss National Science Foundation under grant number 5005-67322.

## 5 References

[1] A. Hasler, I. Talzi, J. Beutel, C. Tschudin, and S. Gruber. Wireless sensor networks in permafrost research - concept, requirements, implementation and challenges. In *Proc. NICOP 2008*, volume 1, pages 669–674, June 2008.

[2] R. Szewczyk, A. Mainwaring, J. Polastre, J. Anderson, and D. Culler. An analysis of a large scale habitat monitoring application. In *Proc. SenSys 2004*, pages 214–226. ACM Press, New York, Nov. 2004.

[3] G. Werner-Allen, K. Lorincz, J. Johnson, J. Lees, and M. Welsh. Fidelity and yield in a volcano monitoring sensor network. In *Proc. OSDI '06*, pages 27–27, Berkeley, CA, 2006

[4] J. Beutel et al. PermaDAQ: A scientific instrument for precision sensing and data recovery in environmental extremes. In *Proc. IPSN '09*, pages 265–276. ACM Press, New York, Apr. 2009.