Predictive Monitoring of Heterogeneous Service-oriented Business Networks: The Transport and Logistics Case

Andreas Metzger*, Rod Franklin†, Yagil Engel‡
*Paluno (The Ruhr Institute for Software Technology), University of Duisburg-Essen, Essen, Germany
Email: andreas.metzger@paluno.uni-due.de
†Kühne + Nagel Management AG, Schindellegi, Switzerland & Kühne Logistics University, Hamburg, Germany
Email: rod.franklin@kuehne-nagel.com
‡IBM Research Labs, Haifa, Israel
Email: yagile@il.ibm.com

***Abstract*—Future service technology will provide an unprecedented access to operational data, which opens up novel opportunities for monitoring, controlling and managing service-oriented business processes. Amongst these opportunities, we consider predictive monitoring to be a major lever for increased efficiency, effectiveness and sustainability in future business networks. Predictive monitoring means that critical events, potential deviations and unplanned situations can be anticipated and proactively managed and mitigated along the execution of business processes.**

This paper demonstrates the potential of predictive monitoring in practice. We focus on transport & logistics as a major industry sector – accounting for between 10% to 20% of a country's Gross Domestic Product. Based on widely adopted standards and real operational data, we empirically support the relevance of key issues faced in that industry sector, such as late cancellations of transport bookings and delayed deliveries. As a solution, we describe the design of a novel, cloud- and services-based collaboration and integration platform. Based on this platform we develop short-term prediction capabilities allowing to proactively manage and mitigate the identified issues in the transport & logistics industry, thus promising to increase business efficiency and sustainability.

***Keywords*-monitoring, prediction, service-oriented architecture, cross-business collaboration, complex event processing, business processes**

I. Motivation

Business processes and networks of the future will be highly digitalized, integrated and interconnected. We will see the seamless integration of ICT services, such as financial and telecommunication services, with physical services, such as transportation or manufacturing [1]. More and more organizations are willing to share information, because they mutually benefit from such information sharing. For instance, such information sharing enables industry benchmarking (e.g., using cross-organizational quality management systems [2]), and it allows for more efficient and effective delivery of products and services (e.g., [3]).

On a technological level, service-orientation will pave the way towards cross-organizational data exchange and integration of IT systems. Ultimately, this integration of heterogeneous and cross-organizational services, together with access to more data sources (e.g., fostered by the availability of cheap sensors), will allow for unprecedented access to operational data from everywhere at any time.

Such data availability and online access to the data opens up opportunities for innovative ways of monitoring, controlling and managing business processes and interactions that build on those services. Amongst these opportunities, we consider *predictive monitoring* (aka. failure / quality prediction [4]–[6]) to be a major lever for increased efficiency, effectiveness and sustainability in future business networks. Predictive monitoring means that critical events, potential deviations and unplanned situations can be anticipated and proactively managed and mitigated [4], [7].

As a simple example, a retailer may have planned the replenishment of its stores on a weekly schedule with each store visited once per week. Weather forecasts (e.g., provided by services accessible over the Internet) provide input to the predictive monitoring system, which may forecast that a weather front will be moving across one of the transport routes. As this leads to a high probability for disrupting scheduled deliveries, the system sends a signal to the retailer's planners to notify them of the potential issue. The planners now have the opportunity to re-plan the scheduled deliveries (e.g., to dispatch an earlier delivery) such as to ensure that the store will not run out of stock due to missed deliveries.

As the example indicates, an industry sector that has the potential to realize significant benefits from predictive monitoring services is the transportation and logistics industry. Transport and logistics activities can account for between 10% to 20% of a country's Gross Domestic Product. Increases in the efficiency of these activities can dramatically improve a country's competitiveness. In addition, environmental impacts resulting from the operation of transport and logistics activities are significant, so any improvement in efficiency within a logistics network may positively contribute to sustainability objectives. As much as 15% of the global greenhouse gas is caused by transportation.

In Section II, we employ the case of transport & logistics to introduce key issues of that sector that may be addressed with predictive monitoring and services technology. Those issues include late cancellations of transport bookings and data delays. In Section III, we then empirically support the relevance of those issues using a widely adopted monitoring standard and real operational data of large forwarding company. In Section IV, we describe the design of a collaboration and integration platform, the FINEST platform, which employs cloud computing and services technology, to deliver predictive capabilities. Section V discussed related work.

II. The Transport & Logistics Case

The transport & logistics industry is one of the largest industries in the world. Consisting of third party service providers as well as in-house service activities, the transport and logistics industry provides freight movement and storage services that allow businesses to source products globally while delivering their goods locally.

Shippers employ air transport when they are concerned about the timeliness of delivery, security, or flexibility in direct shipment to an end delivery point. Because of the high cost of shipping by air, shippers expect exceptional performance in each of these areas. Unfortunately, many factors can contribute to an air cargo shipment not meeting these high levels of expectation.

A. Needs of the Air Transport Industry

The shipment of goods by air requires the integration of services from multiple organizations. Cargo carrying airlines, freight forwarders, airport logistics companies, ground handlers, ground transport companies and regulatory agencies (customs, security, etc.) all may be involved in the physical transport of a good by air. Each of these entities, operating independently to achieve internal efficiency and performance objectives, must be coordinated to act in an integrated fashion if customer expectations are to be met. Delays can arise due to internal considerations such as cargo consolidation, where an organization delays a shipment while waiting to obtain a lower cargo price through a more complete container load. Ground handling delays due to congestion or priority shifts can also arise and contribute to less than desired performance. However, most process oriented delays arise at the boundaries between the entities involved in an air cargo shipment.

Information exchange problems, such as delayed exchange of data, incomplete data exchange or inaccurate data exchange, create significant problems for the timely execution of air shipments. Proper conveyance of information can allow downstream logistics partners to plan for the receipt of goods, link goods receipt to cargo capacity, consolidate goods and plan ground based transport and handling activities. Without this information, downstream portions of the supply chain must react to goods arrivals or, just as importantly, to the non-arrival of previously planned receipts.

B. Key Issues in Air Transport

While there are numerous problems that may occur during transport execution and that would require rescheduling an airfreight shipment, some of the more frequent issues are:

- **Late Shows:** Such issues arise when there is a delay between the expected and the actual time of delivering freight to the carrier (e.g., the air line). Late shows may be due to delays in the transportation from the customer's site to the site of the carrier.
- **No Shows:** Such issues arise if, although a booking has been made by the customer, either no freight is actually available for pickup from that customer, or the booking is canceled late. This means that those issues represent discrepancies between booking and actual.
- **Data Delays:** Where the above issues were implied by deviations and disruptions during the physical transport process, other issues may be caused by the (mis-)alignment of physical and IT processes. Such issues may arise in cases where data is arriving too late into the IT systems; e.g., although cargo has already been loaded into an air plane, the IT systems will know of this only with 30 minutes delay – possibly when the air plane already has departed.

In Section III we provide empirical evidence for the relevance and frequency of such issues. In Section IV, we introduce potential solutions.

III. Empirical Evidence

In this section we provide empirical support for the issues we identified in the previous section by employing data collected following Cargo 2000 [8], a widely adopted process monitoring standard from the air cargo industry.

A. Standard Monitoring Technology

To support the cross-organizational monitoring of transport processes, IATA (the Air Transport Association) has established the Cargo 2000 (C2K) initiative "aiming at implementing a new quality management system for the worldwide air cargo industry. The objective is simple: to implement processes, backed by quality standards, which are measurable to improve the efficiency of air cargo." Cargo 2000 thus enables an unprecedented level of transparency in the supply chain. Specifically, the stakeholders involved in the transport process can share agreed Cargo 2000 messages in an effective and timely manner.

Cargo 2000 is based on the following key principles: (1) Every shipment gets a plan (called a route map) describing predefined monitoring events. (2) Every plan has predefined milestones with estimated time of arrivals. (3) Stakeholders receive alerts (i.e., event notifications) when a milestone has failed.

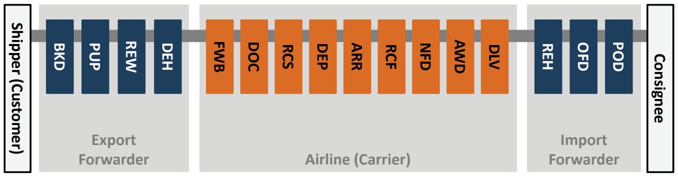

Figure 1. Overview of the Cargo 2000 Process

Figure 1 shows the Cargo 2000 milestones:

- **BKD:** Freight Booking by Shipper[1]
- **PUP:** Pick up from Customer
- **REW:** Freight Received at Export Forwarder[2] Warehouse
- **DEH:** Truck Departure to Airline
- **FWB:** Master Airway Bill[3] Creation
- **DOC:** Truck Arrival at Departure Airline
- **RCS:** Freight Checked in at Departure Airline
- **DEP:** Goods Confirmed on Board
- **ARR:** Flight Arrival at Destination
- **RCF:** Freight Acceptance Arrival Airline
- **NFD:** All Freight and Documents Ready for Pick-up
- **AWD:** Documents Delivery to Import Forwarder
- **DLV:** Freight Delivery to Import Forwarder
- **REH:** Freight Received at Import Forwarder
- **OFD:** Goods Out for Delivery
- **POD:** Delivery of Freight to Consignee[4] & Proof Of Delivery

Air carriers and logistics service providers that have implemented Cargo 2000 are able to track the progress of a shipment by following the various event updates occurring along the shipment route. This is a huge progress from the historical "black box" processes that preceded the industry's implementation of the Cargo 2000 standard.

However, while supply chain participants can see the progress of a shipment, this viewing is in retrospective, i.e., notifications and event updates occur after the fact. Predictive events are not provided through existing implementations of the Cargo 2000 standard. In Section IV we sketch a technical solution to address this gap.

B. Empirical Data Sources

To provide empirical support for our following discussions, we employ real Cargo 2000 monitoring data from a large international forwarding company. More specifically, we use the following three sources of data:

- *Actual Cargo 2000 system messages:* This data set comprises ca. 23,000 Cargo 2000 system messages, containing route map definitions, milestone updates, as well as alerts about violations of milestones.
- *Aggregated Cargo 200 data:* This data set comprises ca. 100,000 actual bookings and transports of the forwarding company for one random month (of the year 2011). Each entry of the data set aggregates information from the related Cargo 2000 messages and includes information such as estimated time of arrival vs. actual arrival, the root cause for delays (if any), as well as additional information about origin and destination of transports.
- *Cargo 2000 quality indicators:* This data set provides industry-wide quality indicators published by IATA [2]. It shows the Cargo 2000 members' overall achievement of KPIs, such as percentage of milestones successfully met.

C. Late shows

Late shows are concerned with potential delays between expected and actual time of delivering the goods to the airline. Late-shows may have different reasons. For example, they may arise due to a delayed pick-up (milestone **PUP**), due to trucks arriving with delay at the forwarder (milestone **REW**), trucks leaving with delay to the departure air line (milestone **DEH**), or delays in checking in the freight at the departure air line (milestone **RCS**).

Analyzing the Cargo 2000 data, there are already close to 1,900 late shows (ca. 2%) which are due to a violation of the milestone **PUP**. This means as much as 63 bookings per day.

Figure 2 shows the distribution of the delays of those late shows. Although we can observe a wide range delays (from one hour to almost 500 hours), over 150 of the late-shows (8%) occur within 10 hours, and 550 (29%) within a 24 hours time frame. Again, an indication that short-term predictions of those events provide an opportunity to take proactive actions.

In order to predict such delays, resorting to Cargo 2000 alone may not be sufficient. Although it may be possible to predict trends based on observing the development and distribution of delays, severe external events may significantly impact on the delays and thus may provide an important source for predictions. As sketched in the introduction to this paper, weather may be one important reason for unforeseen delays and deviations.

As empirical support, Figure 3 shows the rate of **RCS** milestones (freight checked in at air line) violated per month

[1]Shipper, aka. Seller, Exporter, Customer

[2]Forwarder = "An organization which provides logistics services as an intermediary between the shipper and the carrier", cf. http://www.finest-ppp.eu/domain-dictionary

[3]Airway Bill (AWB) = "A shipping document used by the airlines for air freight. It is a contract for carriage that includes carrier conditions of carriage including such items as limits of liability and claims procedures.", cf. ditto

[4]Consignee (aka. Buyer, Importer) = "The person or firm named in a freight contract to whom goods have been shipped or turned over for care.", cf. ditto

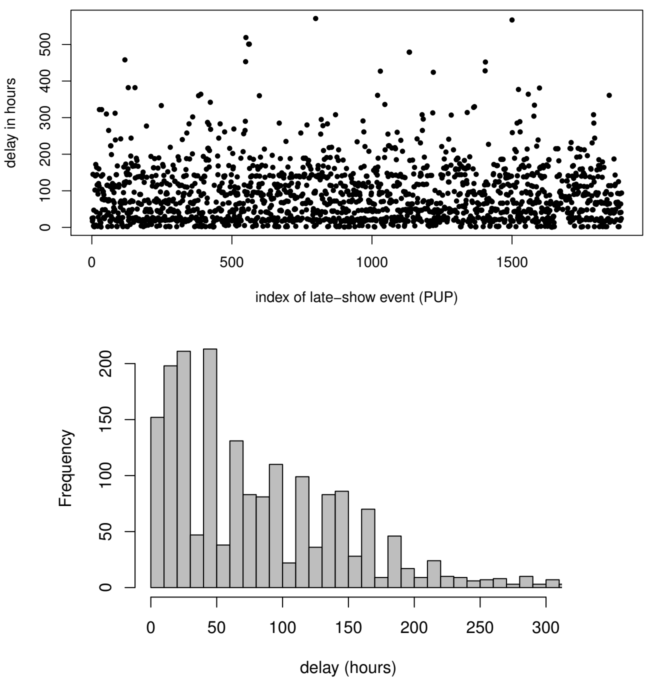

Figure 2. Distribution of delays for late shows (**PUP**)

for a duration of half a year[5]. In April 2010 volcanic ash clouds from an eruption in Iceland visibly impacted on air transport. On a more short-term scale, local, smaller weather conditions may become relevant for anticipating problems.

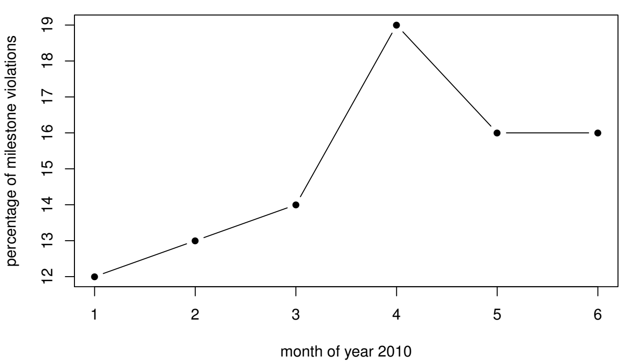

Figure 3. Impact of cumulative late-shows (**RCS**); Data source: [2]

D. No shows

As mentioned above, no shows include situations where although a booking has been made by the customer, (1) no freight is actually available for pickup from that customer, or (2) no freight is actually delivered from the forwarder to the air line. In terms of Cargo 2000 milestones, this means that although milestone **BKD** has been reached, (1) milestone **PUP** or (2) milestone **RCS** fails.

By analyzing the Cargo 2000 data, we were able to determine that – on average for a randomly selected month – as much as 1% of all transport bookings do not lead to an actual pickup (**PUP**). With a rough average of around 100,000 transport bookings of the analyzed month, this means that as much as 1,000 of those bookings may not manifest in an actual transport; that is 33 bookings per day. For the same data set, as much as 0.1% of **RCS** failures have been observed.

Figure 4 shows the distribution of the no shows and late cancellation issues for one month of operational data, focusing on the **PUP** milestone. The diagram shows the number of "successful" transports between these issues. Assuming equal distribution of transport orders over the month, we can conclude that the majority of those issues arise in relatively short intervals. As an example, for an issue with less than 50 "successful" transports in between, the time between the occurrence of two such events will be less than 2 hours. Anticipating those events and acting proactively will thus require short-term prediction capabilities.

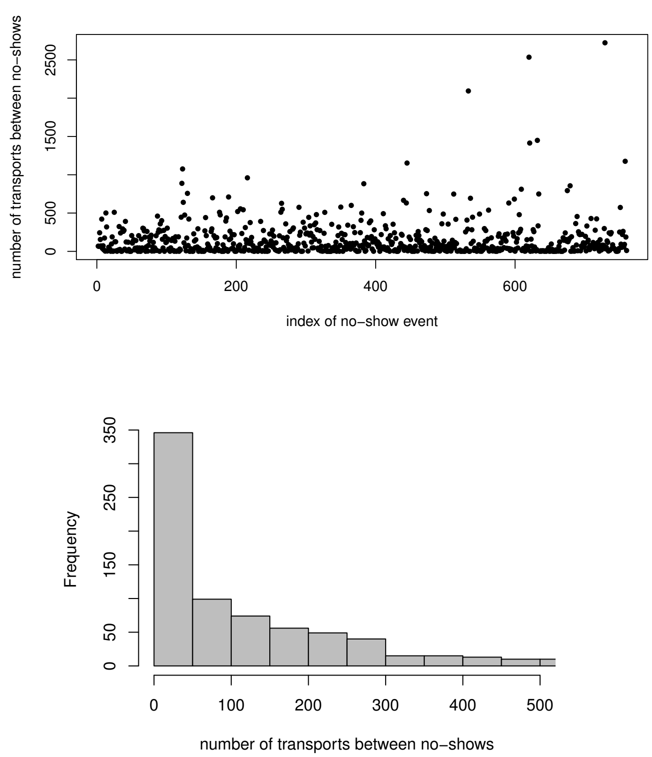

Figure 4. Distribution of no-shows amongst all transports (**PUP**)

E. Data Delays

The above issues arose within the actual physical transport process. However, issues may also arise due to the mis-

[5]Please note that in Figure 3, the delays of milestones previous to **RCS** are accumulated, whereas in Figure 2 we computed the delay only for milestone **PUP** in isolation.

alignment of physical processes and IT systems that support those processes.

One example for such an issue is the delay between the time a physical event occurs (e.g., freight has been physically picked up milestone **PUP**), and the time when such an event is registered in the IT systems. Delays may be due to different factors: (1) Although data is available in real-time in local IT systems, the data might only uploaded periodically and in batch mode to the central IT systems; (2) Data may have been entered too late or even wrong by a human operator.

Issue (1) can obviously be addressed by novel services technology, e.g., such as cloud-based services, which will close the gap between the physical and the digital worlds. Section IV introduces the FINEST platform, which exploits cloud solutions to this end.

Issue (2) appears more difficult to address. Although the Internet of Things may help reduce the gap between physical observation and representation of that observation in the IT systems, as long as humans are involved in that process, delays may remain.

Data delays however, may impact on the ability to predict the behavior of the physical processes. Figure 5 shows the time delays between the registration of a physical event and the availability of this event in the IT systems.

The diagram shows the delays for milestone **FWB** (the creation of the Air Way Bill)[6] As can be seen, most delays (82%) are within 10 minutes. However, some of them (18%) can go up to as much as 60 minutes, a time range, which comes critically close to some of the actual delays in the physical transport process.

Those data delays mean there are situations in which the Air Way Bill physically has been produced at the time defined for its milestone (**FWB**), whereas the IT systems are not aware of that fact. Thus, the IT systems might trigger a false “alarm” indicating the violation of that milestone. Predicting such delays to understand whether there is an actual need for action is very challenging, but promises to alleviate such false “alarms”.

F. Validity Threats

It should be noted that, although we used real operational data and we randomly selected from this data a subset for our analysis, the data was limited to some degree and thus our findings might be limited for what concerns generalization. One important limitation was that the data we employed was not classified according to customers due to privacy reasons. Thus, it may be that the problems we identified concerning no shows are due to a few customers behaving unexpected. As part of our future work, we will thus analyze a larger set of operational data from forwarders and include other data sources in addition to Cargo 2000. Also, we aim to analyze data from other carriers (such as ship or train) to understand whether they face similar issues.

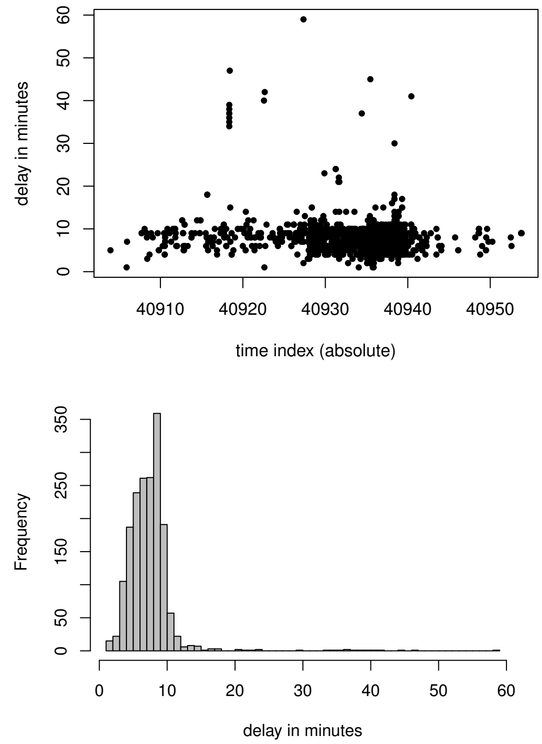

Figure 5. Distribution of data delays measured for milestone **FWB**

IV. THE FINEST SOLUTION

While the transport and logistics industry has made great strides in attempting to improve its efficiency, limitations in technology, transport infrastructure and regulatory regime incompatibilities have created significant barriers to future improvements. ICT technologies promise to overcome these barriers by allowing organizations to rapidly assemble collaborative logistics networks that can efficiently and effectively execute international trading activities.

This section introduces the FINEST platform, a novel ICT solution to overcome those barriers. Following the description of the platform’s design, this section elaborates how the FINEST platform can be employed to address the issues identified in Sections II and III

A. *The* FINEST *Platform*

Building on core capabilities of the Future Internet, currently being developed under the European Union’s Future Internet Public Private Partnership program (FI PPP[7]), the

[6]The data set containing the actual Cargo 2000 system messages was limited to the Airline segment of the T&L process (i.e., milestones **FWB** to **DLV**). We thus chose the first milestone within this segment. Please note that, still today, the Air Way Bill is a physical document (i.e., paper).

[7]Specifically, they are developed within FI-WARE, the European R&D project establishing a Future Internet core platform (see http://www.fi-ware.eu

FINEST platform[8] implements a domain-specific, configurable and extensible set of services for the transport and logistics domain.

As shown in Figure 6, the FINEST platform is structured into three layers. The FINEST platform itself is realized using service-oriented and cloud technology, facilitating interoperability, openness and extensibility through standard interfaces. In addition, the use of integrated security and privacy management mechanisms ensures the secure and reliable exchange of confidential and business-critical information; including mechanisms such as access and identify management, artifact-based security control, and secure storage & backup.

1) Front End Layer: The front end layer of the FINEST platform provides users with role specific, secure, ubiquitous access from different devices to information concerning the operation of the transport and logistics network. The capabilities offered by the FINEST platform and its core modules will be offered through a customizable "web" portal. Each user can configure this portal by selecting dedicated "apps" depending on the capabilities needed to perform a user's respective tasks – quite similar to the iGoogle or iPhone/iPad model. In addition, the front-end will be integrated with messaging systems (such as SMS, E-Mail and Social Networking) such as to notify users and trigger actions.

2) Back End Layer: The back end layer of the FINEST platform provides access to, and integration with, legacy systems, third-party services (Internet of Services) and Internet of Things (IoT) devices. Specifically, the IoT devices will provide (near) real-time information concerning the transport processes as well as their context, thus allowing to quickly and proactively responding. Legacy system integration is facilitated by service-oriented technology, e.g., by exposing features of legacy systems as Web services.

3) Modules Layer: The modules layer of the FINEST platform allows plugging in targeted transport and logistics service modules. Those modules – in the future – may be offered by third parties (such as small and medium enterprises). The initial release of the FINEST platform will feature four open-source modules (called *Core Modules*) for contracting, planning, monitoring and execution of transports:

- *Business Collaboration Module (BCM)* – This module supports the inter-organizational (business-2-business) collaboration between transport and logistics network partners by tracking and tracing transports on the level of business processes and notifying the involved stakeholders in case of deviations or need for action. It may act as an intermediary between these partners and the various cloud based modules selected to manage the efficient flow of goods between the partners.
- *E-Contracting Module (ECM)* – This module provides computer support for service provider selection, contract management and the provision of contract related service requirements to other modules that utilize this information for ensuring the effective and efficient network operation.
- *Transport Planning Module (TPM)* – This module provides support for dynamic transport planning and re-planning activities, exploiting real-time event data provided through the EPM and with respect to contracts between business partners that are managed within the ECM component. Re-planning of shipments occurs when real-time signals from the EPM indicate that a current transport plan cannot be achieved because of some event that has arisen in the shipment process. Such EPM-events will be analyzed by means of the BCM to understand whether re-planning is feasible at all, or whether other actions need to be taken.
- *Event Processing Module (EPM)* – This module provides event-processing facilities to determine relevant situations occurring within and in the context of the transport process. Such events include, for instance, delays of transport (notified from BCM), critical weather conditions (from IoT sensors and IoS services), and violation of Cargo 2000 milestones (from legacy systems). The EPM (based on input from other core modules and the back-end layer) provides the key facilities for predictive monitoring, as explained below.

[8]An initial version has been presented in [9].

B. Addressing Data Delays

One overarching design principle for the FINEST platform was to ensure real-time access and provision of data for all involved stakeholders. The capabilities of the FINEST modules as well as the platform front end will be offered in the Cloud following the Software-as-a-Service (SaaS) delivery model [10]. This will allow stakeholders to directly enter and access operational data. Together with the integration of the Internet-of-Things (by providing service-oriented access sensors, such such as RFID or GPS), this platform design will mitigate data delays in accessing operational information about the transport and logistics processes.

C. Addressing Late Shows and No-Shows

In FINEST, an event-driven architecture is employed for the Event Processing Module (EPM) in order to enable the end-to-end monitoring of a logistics process and to facilitate immediate and proactive response to problems and potential deviations occurring during execution time. An event-driven architecture supports building *reactive* applications, i.e., applications in which processing is triggered by events. This is contrary to traditional *responsive* applications, in which processing is triggered in response to an explicit request from the application [11].

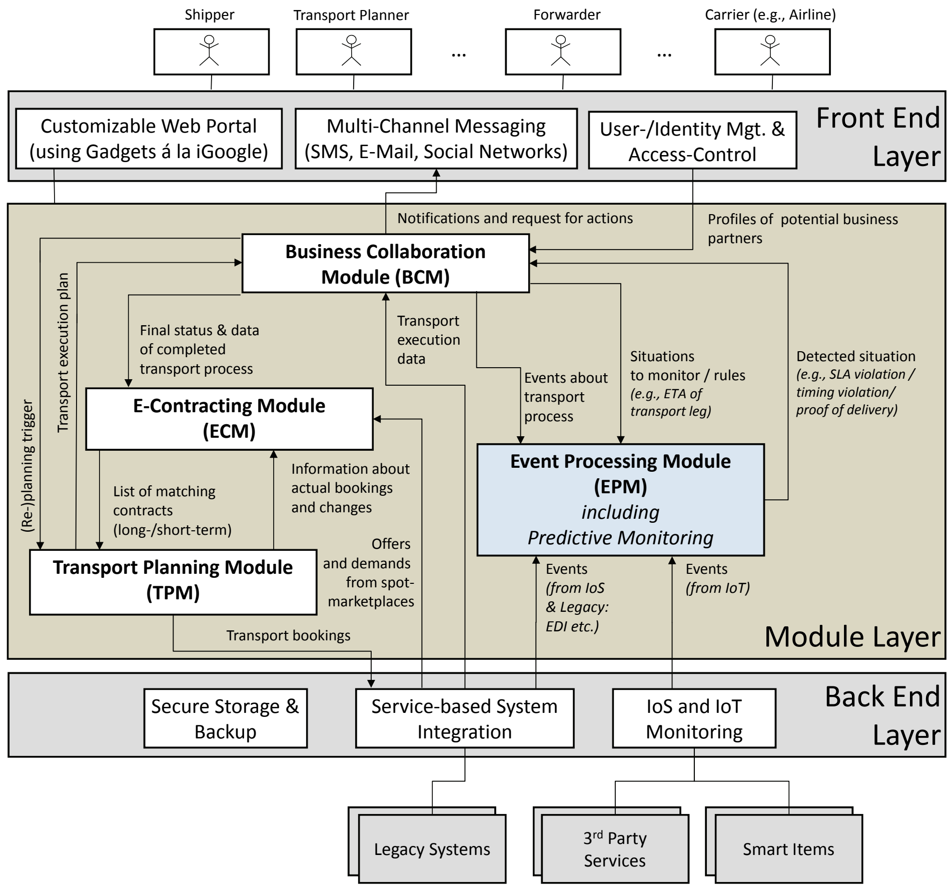

Figure 6. Conceptual Architecture of FINEST Platform

To implement such an event-driven architectural style, FINEST employs Complex Event Processing (CEP), a framework whose purpose is the detection of complex events in streams of incoming raw events. Complex events are characterized as *patterns* of atomic events, comprised of joins, aggregations, filters, and other logical operators. The semantics of a pattern represents a particular situation that is of interest to the system.

The functionality of the EPM can be described at three levels:

- **Track:** Fundamentally, the EPM's event processing capabilities provide visibility into the current status of the logistics process: the location of a shipment, whether it is on a carrier or in a warehouse, whether or not it was customs-cleared, etc.
- **Identify:** Beyond such mere track-and-notify functionality, the EPM defines specific CEP rules that, for example, can detect the *absence* of a milestone achievement on time (e.g., such as the absence of **PUP**, see Section III-D).
- **Predict:** Extending beyond the state-of-the-art capabilities of CEP (see Section V), the EPM leverages the power of event processing towards real-time prediction of potential issues.

Below we provide insights and examples on how the **Predict** functionality of the EPM (together with the capabilities of the FINEST platform in general) helps addressing the late show and no-show issues.

1) Late Shows: As introduced above, late shows occur when there is a delay between the expected and the actual time of delivering freight to the carrier (e.g., the air line). Close and (near) real-time monitoring of the transport process can lead to an early detection of late freight arrival, and this can help to mitigate the impact of the delay through early re-planning (using the TPM module of FINEST); e.g., cargo-space could be used for a different customer, who is predicted to deliver on time.

To this end, CEP rules can compare actual timestamps, provided by real-time milestone update events (such as offered by Cargo 2000 systems), with planned timestamps for the milestones. For an even earlier detection, CEP rules can trigger on the absence of a milestone update within some time frame in which it is expected, assuming that we understand how to handle data delays (cf. Section III-E).

Predictions of the late shows can be calculated from CEP events either analytically (by recalculating the expected time for reaching the following intermediate milestone), or statistically (assuming we found a correlation between late shows and other variables). Furthermore, statistical analysis might establish a correlation between specific exception codes (representing the reason for the delay) and external factors (such as weather conditions). For example, we can accumulate historic data of a customer and find out that this customer tends to timely deliver the freight to the forwarder in only 85% of the cases.

2) No-Shows: As introduced above, no-shows refer to situations in which bookings do not manifest in an actual transport.

Similarly to the aforementioned prediction of late shows, predictions of the no-shows can be calculated from CEP events statistically (assuming we found a correlation between no-shows and other variables). This allows identifying which of the customers tend to cancel their bookings (and in which situations). As an example, accumulating historic data of a customer we may find out that this customer tends to cancel his bookings in 95% of the cases during summer season.

3) Discussion: At first sight the above approaches may seem risky: What if the customer, on a particular occasion, *does* in fact provide the freight on time? One answer to this uncertainty problem may lie in big numbers and careful statistical analysis, allowing to mitigate the risk. A single carrier serves a large number of shippers, a forwarder serves many customers. We thus may apply the predicted amounts rather than booked amounts for all of them, and leave sufficient free cargo space to ensure that the probability of overload is extremely low. Another answer may lie in analyzing the accuracy of the predictions and taking mitigation actions only in cases when those predictions turn out to be accurate enough (see [6] for an discussion on this topic for service-oriented systems).

The above example illustrate that predictive capabilities become particularly useful when integrated with the ability of transport re-planning, such as to mitigate or even avoid the effect of imminent issues. This, in turn, requires real-time collaboration between the various logistic players. The FINEST architecture constitutes an effort of the transport and logistics industry to set up the necessary ICT infrastructure and solutions to allow for such collaboration.

V. RELATED WORK

This section reflects on related work on predicting issues, deviations and failures of service-oriented business processes. To this end, we start with analyzing existing prediction techniques for service-oriented systems and discuss their general characteristics and limitations. We then provide a more detailed discussion of existing work on predictive complex event processing.

A. Predictive Monitoring for Service-oriented Systems

In various areas of computer science and software engineering, predictive monitoring (aka. online failure prediction or online quality prediction) has received considerable attention. A recent survey by Salfner et al. provides an excellent overview and taxonomy of the state of the art in the more traditional area of computer-based systems [4]. Yet, compared with the increasing complexity, dynamics, and flexibility in those more traditional areas [4], service-oriented systems face unprecedented levels of dynamism, together with a lack of control over third-party services [7]. Thus, different classes of novel techniques as well as adaptations of existing techniques have emerged for predicting the quality of service-oriented systems. Below we list the major types of those techniques as identified in S-Cube, the EU Network of Excellence on Software Services and Systems [5].

- *Time series predictors* employ monitoring data (i.e., past observations of service behavior) to extrapolate the future quality of a service; examples include moving averages or exponential smoothing. One important shortcoming of those predictors is that their accuracy may deteriorate in highly variable settings [12].
- *Machine learning/data mining approaches* approaches leverages data mining and machine learning capabilities to train prediction models using historic monitoring data. Shortcomings of these predictors include the fact that they usually required many past observations before providing accurate predictions and need retraining once the system/process has been modified.
- *Run-time Verification* is a formal analysis technique used to ascertain whether some predefined properties are met at run-time; proposed solutions include run-time model checking. Those approaches require a formal model of the system as well as assumptions about how the future execution of the system may develop, which both may be difficult to retrieve in highly dynamic business networks.
- *Static analysis* systematically examines an artifact to infer certain properties. In this regards it is closely related to run-time verification and thus also shares its shortcomings.
- *Simulation approaches* execute dynamic models (in different usage settings) to predict potential future situations. One key shortcoming, why may limit its applicability to the problems address in this paper, is tha executable models (and possibly simulations of the physical environment) are required.
- *Online testing-based approaches* allow complement monitoring data (passively collected) with data actively collected by testing. In settings during which services are less frequently used, this approach allows retrieving additional observations, thus improving prediction ac-

curacy. Yet, in the setting we envision in this paper, limited data is not an issue, and obviously sending cargo around the globe just for "testing" purposed faces significant practical obstacles.
- *Predictive event-processing (CEP)* is able to detect complex events in large streams of incoming raw events, and thus offers clear advantages over the above prediction approaches for what concerns handling "big" volumes of data. We will discuss related efforts in the next section

B. Predictive Complex Event Processing

Event-driven architectures have evolved in recent years, departing from the traditional computing architectures which employ synchronous, request-response interactions between client and servers. Event-driven architectures in general, and complex event processing (CEP) in particular, support reactive applications. This means they enable immediate and automatic response to a set of predefined situations, each characterized up-front during system design and deployment.

While some authors have used CEP in the setting of service-oriented systems (e.g., [13]), they resorted to the traditional use of CEP and have not proposed turning CEP rules into predictive rules. Our work advances from that by employing CEP for the task of *predicting* future events. Previous CEP research did not undertake this direction, with very few exceptions. Among the few, Engel and Etzion [14] introduce *proactive event-driven computing*, a paradigm that combines predictive event-processing with decision making capabilities, targeted at mitigating the effect of predicting undesired events. While their work provides key ideas on the use of CEP for predictive purposes, our work progresses from that by demonstrating how predictive CEP can address practical issues in industry, and how predictive CEP can be integrated into an overall IT solutions for near short-term predictive monitoring.

Considering the issue of data delays, some work related to this problem exists in the CEP field. Several authors have dealt with the more general problem of uncertain events, focusing on the effect that uncertain input events have on the accuracy of the situation that is detected [15]–[17]. These works, however, did not analyze the impact of uncertainty on the prediction of *future* events.

VI. Conclusions and Perspectives

This paper highlighted the potential benefits of exploiting the advancements in ICT, such as the Internet of Things, Cloud computing and Service-oriented Computing, to develop and deploy predictive event monitoring facilities. Using such facilities as part of the FINEST collaboration and integration platform will allow significantly improving the performance and sustainability of transport and logistics processes.

We are confident that the FINEST solutions are also relevant for other application domains. In fact, in the setting of the European Union's Future Internet Public Private Partnership program (FI PPP), application domains such as personal mobility and agriculture indicated they face issues quite similar to those in transport and logistics. As part of our future work, we will thus investigate into extending the scope of the FINEST platform, e.g., by designing additional domain-specific modules.

Acknowledgment

We cordially thank the members of the FINEST project for inspiring discussions and fruitful collaborations; Michael Stollberg for comments on the FINEST architecture and Cyril Alias for stimulating discussions concerning data delays.

The research leading to these results has received funding from the European Union's Seventh Framework Programme FP7/2007-2013 under grant agreements 285598 (FInest), 285248 (FI-WARE) and 215483 (S-Cube).

References

[1] F. Alvarez, F. Cleary, P. Daras, J. Domingue, A. Galis, A. Garcia, A. Gavras, S. Karnouskos, S. Krco, M.-S. Li, V. Lotz, H. Müller, E. Salvadori, A.-M. Sassen, H. Schaffers, B. Stiller, G. Tselentis, P. Turkama, and T. B. Zahariadis, Eds., *The Future Internet - Future Internet Assembly 2012: From Promises to Reality*, ser. Lecture Notes in Computer Science, vol. 7281. Springer, 2012.

[2] IATA, "Mother nature dominates the year in 2010," *benchmark – Newsletter for the supporters of Cargo 2000*, 2011.

[3] N. S. Caswell, C. Nikolaou, J. Sairamesh, M. Bitsaki, G. D. Koutras, and G. Iacovidis, "Estimating value in service systems: A case study of a repair service system," *IBM Systems Journal*, vol. 47, no. 1, pp. 87–100, 2008.

[4] F. Salfner, M. Lenk, and M. Malek, "A survey of online failure prediction methods," *ACM Comput. Surv.*, vol. 42, no. 3, 2010.

[5] A. Metzger, C.-H. Chi, Y. Engel, and A. Marconi, "Research challenges on online service quality prediction for proactive adaptation," in *Proceedings of the ICSE 2012 Workshop on European Software Services and Systems Research – Results and Challenges (S-Cube)*, 2012.

[6] A. Metzger, "Towards accurate failure prediction for the proactive adaptation of service-oriented systems (invited)," in *Proceedings Workshop on Assurances for Self-Adaptive Systems (ASAS), collocated with ESEC 2011*, Szeged, Hungary, 2011.

[7] E. Di Nitto, C. Ghezzi, A. Metzger, M. Papazoglou, and K. Pohl, "A journey to highly dynamic, self-adaptive service-based applications," *Automated Software Eng.*, vol. 15, pp. 313–341, 2008.

[8] R. Cesana, "Cargo 2000 phase 3 specification," IATA/Cargo 2000, Tech. Rep., 2004. [Online]. Available: www.cargo2000.com

[9] R. Franklin, A. Metzger, M. Stollberg, Y. Engel, K. Fjørtoft, R. Fleischhauer, C. Marquezan, and L. S. Ramstad, "Future Internet technology for the future of transport and logistics (invited)," in *ServiceWave 2011, Future Internet PPP Track*, ser. LNCS, A. Zisman, I. Llorente, M. Surridge, W. Abramowicz, and J. Vayssière, Eds. Springer, 2011.

[10] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Commun. ACM*, vol. 53, pp. 50–58, April 2010.

[11] O. Etzion and P. Niblett, *Event Processing in Action*. Manning Publications, 2010.

[12] S. Casolari and M. Colajanni, "Short-term prediction models for server management in internet-based contexts," *Decision Support Systems*, vol. 48, no. 1, pp. 212–223, 2009.

[13] L. Zeng, C. Lingenfelder, H. Lei, and H. Chang, "Event-driven quality of service prediction," *Service-Oriented Computing–ICSOC 2008*, pp. 147–161, 2008.

[14] Y. Engel and O. Etzion, "Towards proactive event-driven computing," in *Proceedings of the 5th ACM international conference on Distributed event-based system*, 2011, pp. 125–136.

[15] S. Wasserkrug, A. Gal, and O. Etzion, "A model for reasoning with uncertain rules in event composition," in *UAI*, 2005.

[16] S. Wasserkrug, A. Gal, O. Etzion, and Y. Turchin, "Efficient processing of uncertain events in rule-based systems," *IEEE Transactions on Knowledge and Data Engineering*, 2010.

[17] S. Wasserkrug, A. Gal, and O. Etzion, "Inference of security hazards from event composition based on incomplete or uncertain information," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 8, pp. 1111–1114, 2008.