Understanding the Quality of Experience in Modern Distributed Interactive Multimedia Applications in Presence of Failures: Metrics and Analysis

Narasimha Raghavan Veeraragavan
University of Oslo, Norway
raghavan@ifi.uio.no

Leonardo Montecchi
University of Firenze, Italy
lmontecchi@unifi.it

Nicola Nostro
University of Firenze, Italy
nnostro@unifi.it

Andrea Bondavalli
University of Firenze, Italy
bondavalli@unifi.it

Roman Vitenberg
University of Oslo, Norway
romanvi@ifi.uio.no

Hein Meling
University of Stavanger, Norway
hein.meling@uis.no

ABSTRACT

Recent technological advances have made it possible to design bandwidth demanding distributed interactive multimedia applications such as the World Opera application. In this application artists from different opera houses across the globe, can participate in a single united performance and interact almost as if they were co-located.

One of the main design challenges in this application domain is to determine the composition of system components necessary to satisfy the desired quality of service in presence of failures and budget constraints. This challenge is exacerbated by the fact that quality of service depends on a multitude of factors such as human perception of video and audio, the type of audience, performance elements, etc. These factors cannot be captured by traditional approaches for dependability evaluation such as reliability, i.e., continuous delivery of correct service. This calls for developing a more comprehensive "Quality of Experience" concept.

In this paper, we propose a novel method to assess the quality of experience in presence of failures, based on a new metric called perceived reliability. This method can help the system designers and engineers compare architectural variants and to determine the dependability budget. We show the feasibility of our method by applying it to a World Opera performance. Our experimental results provide useful guidelines for system engineers towards improving the quality of experience of World Opera performances despite presence of failures.

Categories and Subject Descriptors

C.2.4 [**Distributed Systems**]: Distributed applications;
C.4 [**Performance of Systems**]: Modeling techniques, fault tolerance, reliability, availability, and serviceability

Keywords

Reliability analysis, Quality of Experience, Distributed Interactive Multimedia Applications, World Opera

1. INTRODUCTION

Distributed Interactive Multimedia Applications (DIMAs), such as Massive Online Multiplayer Games [9] and video conferencing systems have become commonplace among users of the Internet. These applications share a number of demanding traits, such as their real-time and interactive nature, which imposes stringent requirements on latency and synchronization. These classical DIMAs are limited in terms of their interaction complexity, types and number of streams, and number of participants at each location. Recent technological advances have enabled the design and implementation of even more demanding DIMAs, in which the bandwidth requirement vastly surpasses that of classical DIMAs. One such application is World Opera (WO), an application envisioned by the WO artistic consortium [3].

The WO consortium and its partners are engaged in conducting distributed, real-time, live opera performances across several world renowned opera houses. Each opera house represents a real-world stage with its own musicians, singers, dancers, and actors. Interaction between the artists is orchestrated by a single conductor present at a single selected stage. Participating artists from different real-world stages map to *virtual-world stages*, which are projected as video on display devices, and shown to the audience at the local opera house as well as audiences at geographically distributed (remote) opera houses. Additionally, virtual-world stages can display animated cartoon characters mimicking the behavior of the artists at remote stages. The virtual-world and real-world stages together form a mixed-reality stage. A collection of distributed mixed-reality stages together constitute a WO application.

The pictures in Figure 1 shows WO rehearsal experiments conducted at the three different rooms at the Music Conservatory of the University of Tromsø. The first experiment (shown in Figure 1a) includes a singer with a remote pianist and a remote conductor in a large black-box theatre; the second scenario has three performers located in three different locations (shown in Figure 1b and in Figure 1c). More details about the rehearsals are available in [3]. A key observation derived from the early experimentations with WO performances reveals that it is notoriously difficult to maintain a smooth technical operation for the entire duration of a performance,

(a) Singer listens to the remote conductor

(b) Singer coordinates with a remote pianist and a remote actor

(c) Actor coordinates with a remote singer and a remote pianist

Figure 1: World Opera Experiments conducted at Music Conservatory at University of Tromsø

even after taking proper preparatory steps.

Towards addressing this challenge, we observe that, while it might be difficult to deliver a flawless performance in DIMA applications such as WO, these applications are characterized by the possibility to define meaningful graceful degradation. For example, it is still acceptable for the audience to hear the orchestra for a moderate duration of time without seeing it. Secondly, we advocate that subjective factors, including the perception and other characteristics of users, have a key role in correctly evaluating the reliability of multimedia applications. A completely failure-free execution is not necessarily required to accomplish a successful performance. For example, a microphone outage of ten milliseconds is not discernible by the human audience; on the other hand, the failure of one microphone from a group of instrumentalists may be tolerated by the audience for a limited amount of time. Therefore, the classical notion of reliability (*continuous* delivery of correct service) is not an appropriate metric for evaluating DIMA applications in which partial and intermittent failures may not necessarily impair the performance. To this end, we propose a novel reliability metric taking into account user-perceived *quality of experience* (QoE)[18]. This metric, called *perceived reliability*, is based on the concept of *tolerable failure duration*, which limits the duration of intermittent failure intervals. Finally, we provide a modeling framework that allows such metric to be evaluated, thus facilitating decisions about the design and setup for WO performances.

In summary, we provide the following key contributions in this paper: i) we apply a QoS approach to capture the concept of meaningful degradation in DIMA applications; ii) we define QoE-aware metrics which take into account subjective perception of users; iii) we design and implement a modeling framework to evaluate such metrics; and iv) we apply the framework to a WO performance, providing useful guidelines to stage engineers.

2. WORLD OPERA ARCHITECTURE

The typical setup for a World Opera performance consists of 3 to 7 real-world stages with one stage acting as a main stage, where the conductor and a large number of artists are physically located. Additionally, the main stage has more complex and a large sets of technical components compared to all other stages. The activities at each stage in WO are divided into five phases as shown in Figure 2: initialization, capturing, processing, streaming, and rendering. During the *initialization phase*, the directors of all stages agree on the set of components that need to be used on each stage for the session. In the *capturing phase*, the corresponding components receive activation signals and start generating streams. There exist three principal stream types: video, audio and sensor (e.g., to track the movement of an artist on the stage). These generated

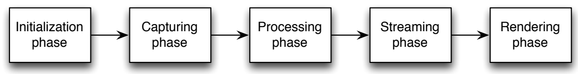


Figure 2: Phases of operation in World Opera

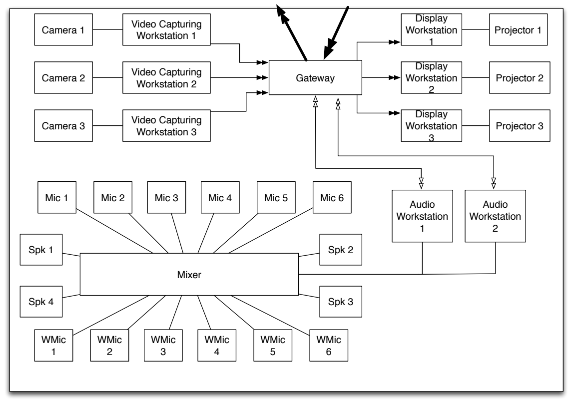


Figure 3: System architecture of a World Opera stage

streams collectively represent the real-world data.

In the *processing phase*, all generated streams are processed to remove noise. Additionally, video streams are encoded to reduce the size of streams, timestamped, and processed using computer vision techniques for artistic reasons. The *streaming phase* is where the streams are multicast and received by the remote stages. In the *rendering phase*, the received streams are processed (e.g., decoded), synchronized based on their timestamps and then rendered to the virtual-world.

The architecture considered for the main stage of WO is shown in Figure 3. In order to cope with faults that may affect stage components, most of the components have hot standby spares. This design choice is due to the fact that the repair of a component is considered too time consuming to be practical during a WO performance, since it would severely impact the show. Therefore, all the components are non-repairable during the performance: as soon as a component is failed, its functionality can be restored by switching to an identical spare.

In our analysis we focus on the main stage as all other stages have relatively less, simple and reliable components. The same approach can be however be extended to minor stages as well. Additionally, we ignore the initialization phase because it is performed offline before the performance starts. Failures in this phase do not significantly impact the reliability of the online performance since the repair rate in this phase is high; i.e. there is a high probability that failed components are replaced before leaving the initialization phase. Note that the processing phase and the sensor streams are not part of the current WO deployment, and thus for simplicity in this paper we also exclude them from our models.

We now present the principal sets of components that are employed in the capturing, streaming, and rendering phases. The **capturing phase** involves the following components: cameras, camera workstations, wired and wireless microphones, and a mixer. Camera workstation includes both hardware and software while the other components are hardware-only. Cameras are used to capture the video streams portraying the artists from multiple viewpoints in the real-world. The captured streams are sent to the camera workstations, a set of enterprise strength computers equipped with a high-speed network interface card. Camera workstations are used to receive generated streams from the cameras and control the motion of the cameras during the performance. Multiple camera workstations send their received video streams to the gateway.

Wired and wireless microphones are organized into a microphone array controlled by the mixer. The mixer is responsible for activation of microphones, adjustment of audio signals, and routing the audio signal. Microphones generate audio streams representing the sound produced by artists in the real-world. The generated audio streams are routed to multiple speakers via the mixer, and to the remote stages through the gateway via audio workstations.

The **streaming phase** involves a single hardware component: the gateway. The gateway is responsible for multicasting the video and audio streams to the remote stages through a dedicated high-speed connection of at least 10 Gbps. These streams are received at the remote stages and forwarded to the display and audio workstations.

The **rendering phase** has the following components: display and audio workstations, projectors, mixer, and speakers. Workstations include both hardware and software while the other components are hardware-only. The audio and video streams are received by the audio and display workstations respectively. The audio workstations render the audio streams to the speakers with the aid of a mixer. The display workstations render the images to multiple screens using the projectors.

3. PERFORMABILITY MODEL

3.1 Modeling assumptions and notation

In this work we aim at evaluating and comparing the performability properties offered by different setups of WO stages. The main objective is to support stage technicians in ensuring an acceptable quality level for the audience, possibly using more reliable components or adding spare elements to the architecture. The perfomability model described in this section analyzes the main stage S_i of the WO, which is assumed to be composed of 10 types of components: wired microphones (*mic*), wireless microphones (*wlmic*), cameras (*cam*), projectors (*proj*), speakers (*spk*) display workstations (*ws_disp*), audio workstations (*ws_aud*), camera workstations (*ws_cam*), mixers (*mix*) and gateways (*gw*).

The model is based on the following set of assumptions:

- Stage components are only subject to crash failures: when a component fails it completely stops working and produces no output. Failures that cause components to produce incorrect output (e.g., noise) are not considered. Failures of components of type *class* are exponentially distributed with rate λ_{class}.
- Switching takes an exponentially distributed amount of time with mean τ_{class}, which includes the time needed to detect the failure of a component. Switching fails with probability $1 - c_{class}$, resulting in the spare component becoming unusable for the rest of the performance.
- When a workstation has failed, the functionality of all its connected components (e.g., cameras) is lost. If a spare is successfully activated for the workstation their functionality is restored (unless they have failed as well).

3.2 Measures of interest

The objective of our analysis is to establish a framework that allows us to evaluate how the physical setup of the stage affects the Quality of Experience (QoE) of the planned WO performance. QoE is a broader concept than QoS, adding important properties unique to DIMAs, such as human perception of video and audio.

Being able to evaluate if (and to what extent) the system fulfills the expectations of different classes of audience is of primary importance in multimedia applications, and even more in distributed multimedia applications like the WO, which are potentially affected by multiple sources of quality degradation. It should be noted that the offered QoE of a WO application is influenced by several factors, including the content and characteristics of the performance. For example, small interruptions of audio streams could be hardly noticeable during slow monologues, while they would have significant impact for fast paced music.

One of the main aspects that should be addressed in the evaluation of the QoE provided by the WO is the ability of the system to provide an acceptable performance quality even in presence of failures. What "acceptable" means is clearly subjective and it depends on multiple factors. From a QoS perspective, several *quality levels* can be defined for the system, based on the objective level of degradation of the offered service, enabling the evaluation of *performability* [13] offered by the system.

The WO architecture defined in Section 2 is composed of two main subsystems: the *audio* subsystem, comprising microphones, speakers, the mixer and audio workstations, and the *video* subsystem, comprising cameras, projectors, and related workstations. In such a system, one of the most natural ways to define different quality levels is to consider the two subsystems separately, requiring both of them to be fully functional for full system functionality, but allowing degraded modes where only one of the subsystems is correctly working. The concept of tolerable degradation depends on the performance type: for example, quality levels in which the video is degraded could be tolerated during concerts, while they would be much more disruptive during a dance performance.

In principle, several quality levels can be defined based on the kind of performance, stage setup, and artists lineup. A simple way to define different quality levels for the system is based on the number of components that are correctly working on the stage. We define the quantity N_{class} as the total number of primary[1] components of type *class* defined by the stage architecture and w_{class} as the number of primary components of type *class* that, at a given instant of time, are correctly performing their function.

Based on these quantities, we can define a set of n quality levels for the audio subsystem, $\{A_1, \ldots, A_n\}$, and m levels for the video subsystem, $\{V_1, \ldots, V_m\}$, leading to $n \times m$ combinations for the overall performance. In this paper we focus on the definition of the modeling approach, rather than on defining all meaningful quality levels for a real performance; therefore a limited set of quality levels is considered.

Table 1 defines three levels for the audio subsystem, based on the number of primary components from different classes that are correctly performing their function at a given instant of time:

- A_1 represents the highest quality level, in which all the audio subsystem components are working correctly.
- A_2 represents a slightly degraded delivery of the performance,

[1]Primary components are the default components of the architecture without any spares.

in which the number of failed microphones or speakers is less than or equal to the threshold value. The threshold values γ_{class}, with $0 < \gamma_{class} < N_{class}$, determine what configurations might still be acceptable considering the kind of performance and characteristics of the audience. For example, if two or three out of ten microphones used for an orchestra have failed, this results in a slightly degraded QoE.

- A_3 represents the worst (and likely, unacceptable) quality level. For example, if more than three microphones used for an orchestra have failed, then the audience perceives a significant degradation.

Table 1: Quality levels for the audio subsystem

A_1	$(w_{mic} = N_{mic}) \wedge (w_{wlmic} = N_{wlmic}) \wedge (w_{spk} = N_{spk}) \wedge (w_{gw} = N_{gw}) \wedge (w_{mix} = N_{mix})$
A_2	$(\gamma_{mic} \leq w_{mic} \leq N_{mic}) \wedge (\gamma_{wlmic} \leq w_{wlmic} \leq N_{wlmic}) \wedge (\gamma_{spk} \leq w_{spk} \leq N_{spk}) \wedge (w_{gw} = N_{gw}) \wedge (w_{mix} = N_{mix}) \wedge \neg A_1$
A_3	$\neg A_1 \wedge \neg A_2$

Table 2: Quality levels for the video subsystem

V_1	$(w_{cam} = N_{cam}) \wedge (w_{proj} = N_{proj}) \wedge (w_{gw} = N_{gw})$
V_2	$(\gamma_{cam} \leq w_{cam} \leq N_{cam}) \wedge (\gamma_{proj} \leq w_{proj} \leq N_{proj}) \wedge \wedge (w_{gw} = N_{gw}) \wedge \neg V_1$
V_3	$\neg V_1 \wedge \neg V_2$

Similar to Table 1, Table 2 describes three quality levels for the video subsystem. The quality levels for the overall system can then be derived as:

$$Q_{\alpha\beta} = A_\alpha \wedge V_\beta,$$

ranging from level Q_{11} representing "perfect" operational conditions, to Q_{33}, which represents the lowest quality level. It should be noted that workstations are not explicitly considered for determining the quality level provided by the WO; however, since their failure causes the loss of the functionality of components connected to them, they are implicitly taken into account in the evaluation.

The time spent in each of these levels, $\mathcal{T}_{\alpha\beta}$, provides a system-level, QoS-oriented, view of the quality of the delivered performance. However, for the purposes of QoE evaluation, user-oriented measures are of greater interest. As a first step, we identify a minimum quality level, $Q_{\alpha\beta}$, that the user is willing to accept during the WO performance. Then, we define the *availability* of the show in the interval $[0, t]$ with respect to $Q_{\alpha\beta}$, denoted as $\mathcal{A}_{\alpha\beta}(0, t)$, as the fraction of time in which the system is providing a quality level Q_{ij} such that $i \leq \alpha$ and $j \leq \beta$.

Following its classic definition [4], the *reliability* of a WO performance with respect to $Q_{\alpha\beta}$, denoted as $\mathcal{R}_{\alpha\beta}(t)$, is the probability that, up to time t, the system has been continuously providing a quality level equal to or better than $Q_{\alpha\beta}$.

While this definition is a useful system-level measure, we find it too restrictive for multimedia applications like WO; in fact, it does not take into account the actual perception of the user, who may not be able to notice quality degradation for a limited amount of time or may be able to tolerate it, if a higher quality level is restored within a reasonable amount of time. For this reason, for each quality level A_i and V_i, with $i > 1$, we define the *tolerable duration*, $\vartheta(A_i)$ (equivalently, $\vartheta(V_i)$), as the maximum amount of time that the user may tolerate such quality level without considering the whole performance affected. Moreover, $d_{ij}^{max}(t)$ is the duration of the longest time interval in $[0, t]$ in which the system has been *continuously* delivering quality level Q_{ij}.

We then define the *perceived reliability* of a WO performance with respect to $Q_{\alpha\beta}$, denoted as $\hat{\mathcal{R}}_{\alpha\beta}(t)$, as the probability that:

$$d_{ij}^{max}(t) \leq \min\left\{\vartheta(A_i), \vartheta(V_j)\right\}, \qquad \forall i > \alpha, j > \beta,$$

i.e., the probability that, up to time t, the system has not been providing quality levels lower than $Q_{\alpha\beta}$ for a duration that is not "tolerable" by the audience.

3.3 Stochastic Activity Network Model

To evaluate such metrics and compare the QoE perceived by different users under different stage configurations, a Stochastic Activity Networks (SANs) model has been constructed. SANs [16] can be considered an extension of the well-known Stochastic Petri Nets (SPNs) [5] formalism; they provide a greater flexibility by introducing additional primitives and the possibility to use non-exponential transition delays. SANs are supported by the Möbius modeling framework [7], which provides both numerical and simulation-based methods for their evaluation.

In building the model, a modular and compositional approach is employed, allowing different stage configurations to be easily modeled and evaluated. A set of "template" SAN atomic models have been created, each representing one of the eight component classes; atomic models are then assembled using the Join/Rep composition formalism [15] to obtain the overall model of the WO stage. Due to space limitations, we only provide a high level overview of the composed model in this paper without discussing in-depth details of each atomic model. Further details on our SAN model construction can be found in the technical report [14].

Table 3 briefly describes the atomic models (corresponding to the leaves in Figure 4) that are identified as building blocks for generating the overall stage model. Figure 4 shows how these atomic models are composed and replicated in order to obtain the model representing the overall stage. The model `JoinCameraWS` represents the video capturing phase. This model is obtained by the join of the camera workstation and the `RepCameras` corresponding to all the cameras connected to that workstation. This model is then replicated to obtain all the necessary components of the stage.

The model `JoinProjectorWS` represents the video rendering phase. The model is obtained by the join of the display workstation and the `RepProjectors` corresponding to all the projectors connected to that workstation. This model is then replicated to obtain all the necessary components of the stage.

The model `JoinMicsWS1` represents the audio capturing and rendering phase. The model is obtained by joining the audio worksta-

Table 3: Description of the atomic models

`gateway`	gateway responsible for multicasting the streams.
`mixer`	mixer that is responsible for the audio components.
`ws_camera`	generic camera workstation associated to a group of cameras.
`camera`	generic camera component.
`ws_display`	generic display workstation associated to a group of projectors.
`projector`	generic projector component.
`speaker`	generic speaker component.
`ws_audio1`	audio workstation associated to a group of speakers and wired microphone.
`microphone`	generic wired microphone.
`ws_audio2`	audio workstation associated to a group of wireless microphone.
`microphone_wless`	generic wireless microphone.
`component_counter`	atomic model used as support for the definition of metrics to be evaluated.

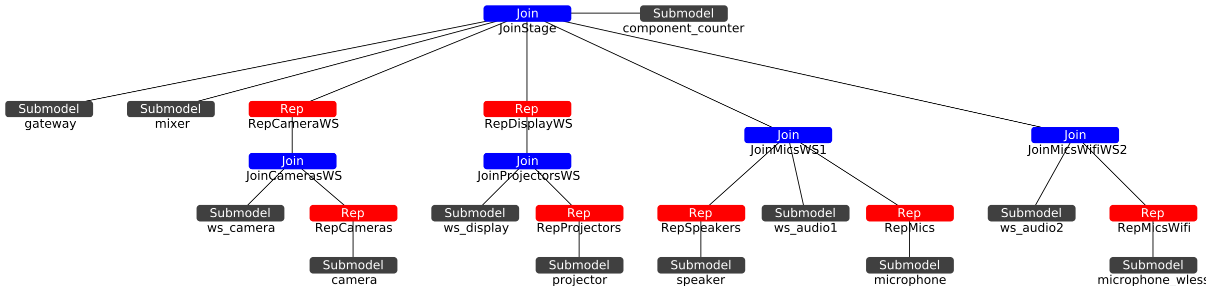


Figure 4: Composed model of a generic stage of the WO performance

tion and the replicas of the speaker and wired microphone connected to that workstation. This model is not replicated because of the architecture of a WO stage described in Section 2, which means that only one audio workstation is involved to work with speakers and wired microphones. The model `JoinMicsWifiWS2` represents the audio capturing phase related to the wireless microphones. This model is obtained by joining the audio workstation and the replicas of wireless microphone connected to that workstation. Finally, all such models are joined to form the root `JoinStage` model, which represents the overall stage under study. To support the definition of the target metrics, the additional atomic model `component_counter` is added to the topmost join node.

4. EVALUATIONS AND RESULTS

The model described in Section 3 has been evaluated using the discrete-event simulator provided with the Möbius framework [7]. Analytical solution was not feasible, since our model allows more than one deterministic activity to be enabled at a time [5]; this is necessary to correctly represent the elapsing of tolerable duration intervals. Moreover, even if considering exponential activities only, the state-space explosion problem arises as soon as more spares are added to the stage architecture.

The main objective of our experiments is to evaluate the perceived reliability wrt various QoE levels. Additionally, we evaluate the effect of introducing, comparing the impact of adding spares to less reliable components only, or to all components of the architecture. Furthermore, we analyze the consequences of reducing the switching time, varying the tolerable duration, and varying the failure rates of the less reliable components. Moreover, we evaluate the effect of adding spares on the perceived reliability of various QoE levels. Finally, we compare the traditional reliability with the perceived reliability.

4.1 The analyzed scenario

The evaluations analyze the quality of the performance delivered at a single WO stage, based on the stage architecture described in Section 2. Table 4 summarizes the main model parameters and their default values that have been used in the following evaluations. Model parameters are divided in two groups: stage parameters, describing the stage architecture, and audience parameters, characterizing the audience. For each component type, stage parameters describe its failure rate λ, the probability of successful switch c, the average time it takes to switch to a spare component τ, and the threshold γ that is considered for the definition of quality levels. Audience parameters define the tolerable durations ϑ that are associated with degraded quality levels.

The failure rates shown in Table 4 are conservative estimates based on the literature [17]. The coverage probability, switching time, and the threshold defined for quality levels are also our conservative estimates based on the discussion with the technicians. While our model is independent of the failure distribution, we assumed for our experiments that failure rates are exponentially distributed following conventional practice. For all experiments, we fix the mission-time to 2 hours, representative of most Opera performances.

Table 4: Main model parameters and their default values

Stage parameters					
Component type	λ ($hours^{-1}$)	N	c	τ (secs)	γ
Camera	0.002	3	0.95	60	50%
Camera workstations	0.00001	3	0.95	180	-
Display workstations	0.00001	3	0.95	180	-
Audio workstation 1	0.00001	1	0.95	5	-
Audio workstation 2	0.00001	1	0.95	5	-
Projector	0.006	3	0.95	60	50%
Speaker	0.001	3	0.95	1	50%
Wired microphone	0.002	4	0.95	5	50%
Wireless microphone	0.002	4	0.95	120	50%
Mixer	0.0001	1	0.95	5	-
Gateway	0.00001	1	0.95	5	-

Audience parameters			
	Tolerable duration ϑ		Tolerable duration ϑ
Level A_2	30 seconds	Level V_2	30 seconds
Level A_3	1 second	Level V_3	1 second

4.2 Results

The amount of time spent in each quality level during the performance ($\mathcal{T}_{\alpha\beta}$) is shown in Table 5 for different number of spares. Since results have been obtained by discrete-event simulation, for each value the confidence interval is shown as well; it should be noted that values smaller than 10^{-8} corresponds to an amount of time smaller than 1 ms, which would be hardly noticeable by the audience. The amount of time spent providing the highest quality level is considerably increased by adding spare components; adding spares for all components provides a slight increase with respect to adding spares for the least reliable components only (i.e., microphones, cameras, projectors, and speakers).

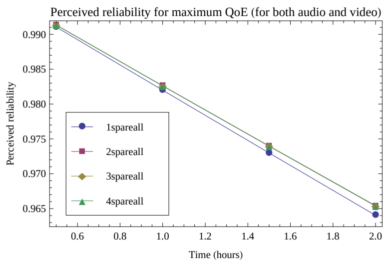
4.2.1 Perceived reliability with respect to different QoE levels

Figure 5a shows the impact of adding spares on the perceived reliability with respect to the highest quality level, $\mathcal{R}_{11}(t)$. Adding a second spare to all the stage components still improves the probability to deliver a "perfect" performance to the audience; adding three or more spares only produce limited improvements, and therefore might not be worth the extra costs. The same observation can also be obtained from the analysis of Figures 5b, 5c, and 5d, which show the perceived reliability with respect to the quality levels Q_{13} (perfect audio), Q_{31} (perfect video), and Q_{22}, respectively.

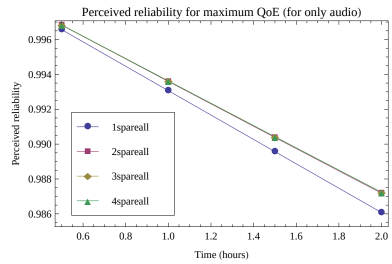
The results in Figure 5c show that adding spares provides only a minimal improvement with respect to the quality of the video subsystem only. This is caused by the higher time required to switch to spares for some components of the video subsystem, camera and

Table 5: Effect of introducing spares on time spent in various levels of QoE ($\mathcal{T}_{\alpha\beta}$) in a 2 hours performance (values in hours)

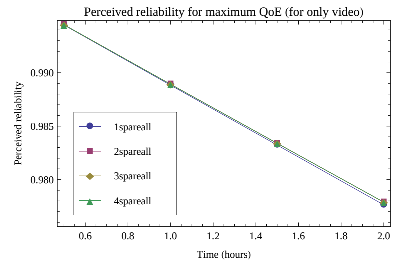
Levels	No spares to all components	1 spare for all components	2 spares for all components	1 spare – less reliable only	2 spares – less reliable only
A1V1	1.916 ± 1.977E-05	1.994 ± 1.39E-05	1.998 ± 3.56E-06	1.993 ± 1.41E-05	1.997 ± 5.223E-06
A1V2	3.44E-02 ± 1.28E-05	2.62E-03 ± 9.21E-06	7.43E-04 ± 2.35E-06	2.68E-03 ± 9.12E-06	7.98E-04 ± 2.76E-06
A1V3	3.07E-04 ± 1.06E-06	1.57E-06 ± 1.81E-07	1.03E-07 ± 8.63E-09	1.73E-06 ± 1.87E-07	1.18E-07 ± 1.38E-08
A2V1	3.62E-02 ± 1.31E-05	2.49E-03 ± 9.13E-06	6.80E-04 ± 2.34E-06	2.49E-03 ± 8.89E-06	6.81E-04 ± 2.28E-06
A2V2	8.77E-04 ± 1.80E-06	3.95E-06 ± 2.93E-07	2.54E-07 ± 2.35E-08	4.18E-06 ± 2.94E-07	2.88E-07 ± 2.71E-08
A2V3	8.87E-06 ± 1.62E-07	7.32E-10 ± 5.59E-10	5.14E-11 ± 1.21E-10	9.53E-10 ± 1.31E-09	4.32E-12 ± 1.11E-11
A3V1	2.31E-04 ± 1.06E-06	1.06E-05 ± 6.61E-07	8.20E-07 ± 1.44E-07	2.39E-04 ± 3.10E-06	2.39E-04 ± 3.10E-06
A3V2	5.69E-06 ± 1.45E-07	2.59E-08 ± 2.63E-08	1.74E-10 ± 1.78E-10	4.23E-07 ± 1.03E-07	1.07E-07 ± 3.03E-08
A3V3	1.98E-05 ± 3.11E-07	9.91E-07 ± 1.96E-07	8.67E-08 ± 4.73E-08	1.90E-05 ± 8.75E-07	1.93E-05 ± 8.86E-07



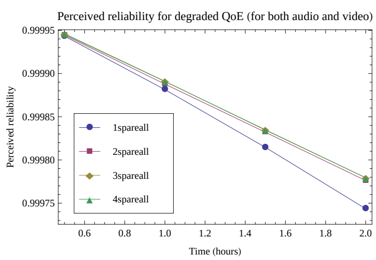
(a) $\hat{\mathcal{R}}_{11}$ – adding spares for all components



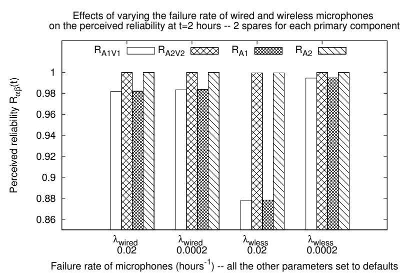
(b) $\hat{\mathcal{R}}_{13}$ – adding spares for all components



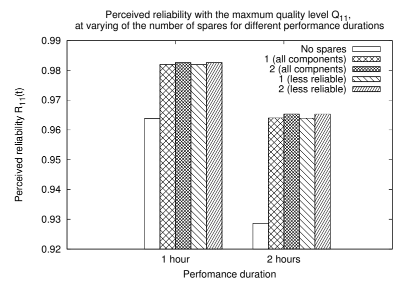
(c) $\hat{\mathcal{R}}_{31}$ – adding spares for all components



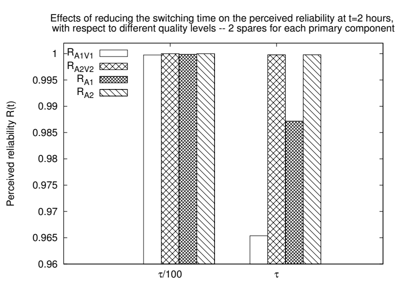
(d) $\hat{\mathcal{R}}_{22}$ – adding spares for all components



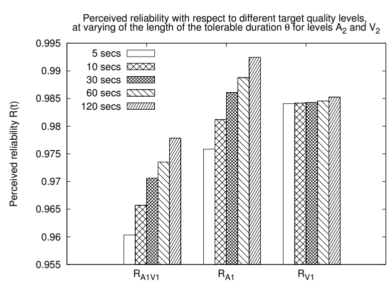
(e) Varying the failure rate of microphones



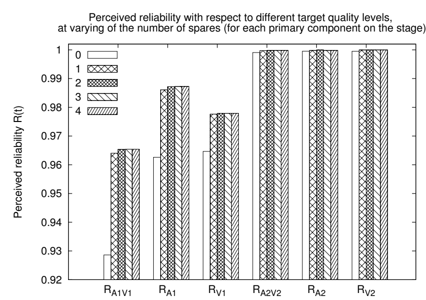
(f) Adding spares to all components, or just to less reliable ones



(g) Reducing the time required to switch to spare components



(h) Varying the "tolerable duration" ϑ parameter



(i) Effect of adding spares on different quality levels

Figure 5: Perceived reliability $\hat{\mathcal{R}}(t)$ for different stage configurations and with respect to different quality levels

display workstation in particular. Since the time required to perform the switching is high, there is a high probability that it exceeds the tolerable duration. In contrast, significant improvement in the audio QoE is perceived by the audience when more spares are added (Figure 5b), since most components of the audio subsystem have a small switching delay.

4.2.2 Comparing architectural variants

Figure 5e considers a stage with two spares for each primary component, and it shows how the failure rate of microphones impacts the perceived reliability of the show. Results show that, in this setup, the failure rate of *wired* microphones does not affect the overall performance, since two spares are available and the switching delays of such components is rather low (5 seconds). Conversely, the failure rate of *wireless* microphones heavily affects the performance, and employing more reliable wireless components helps in increasing the perceived reliability with respect to QoE levels involving the audio subsystem.

If deciding to add spares to the stage, it may be more convenient, for budget reasons, to provide spares only for certain component types. In particular, it can be easy to find spare elements for less expensive (and less reliable) components like microphones or camera, but it may be more difficult and more expensive to retrieve spares for other components like the mixer or the gateway. Figure 5f shows how the reliability for the maximum quality level ($\mathcal{R}_{11}(t)$) is improved by adding spares to all the components, or by adding spares just to less reliable ones, i.e., microphones, cameras, speakers, and projectors. Results for both 1 hour and 2 hours of performance are drawn in the figure, both showing that adding spares to all components does not provide a significant improvement.

Figure 5g shows the impact of switching delays τ on the perceived reliability with respect to different quality levels. Two stage configurations are compared in the figure: one in which default τ values from Table 4 are used, and one where such values are reduced by two orders of magnitude (i.e., $\tau' = \tau/100$). Such a great improvement in switching delay could be achieved by introducing automated mechanisms for failure detection and failover, which could justify, for example, a decreasing of microphones switching times from 60 seconds to 600 ms. Decreasing switching times greatly improves the perceived reliability, since there is a higher probability that in case of failure a spare is activated without the audience even perceiving an interruption of the show.

4.2.3 Comparing solutions based on the target QoE and performance type

In Figure 5h it is analyzed how the perceived reliability with respect to different levels is affected by the tolerable duration $\vartheta(A_2)$ and $\vartheta(V_2)$. Five different configurations are considered, ranging from an audience which is able to tolerate just 5 seconds of degraded audio and video quality, to an audience that is capable to tolerate up to 2 minutes of degraded quality. Results show that, in general, the perceived reliability increases with the increase of the amount of time that the audience (or the kind of performance) is able to tolerate. However, the perceived reliability with respect to the video subsystem only ($\mathcal{R}_{31}(t)$) remains almost unchanged: only a little improvement is noticeable when $\vartheta(V_2)$ is above 60 seconds. It is interesting to note that when the tolerable duration is low, the perceived reliability with respect to the video subsystem only (i.e., $\mathcal{R}_{31}(t)$) is higher than that with respect to the audio subsystem only (i.e., $\mathcal{R}_{13}(t)$). However, as $\vartheta(A_2)$ and $\vartheta(V_2)$ increase, the perceived reliability with respect to audio increases, while that with respect to video remains almost unchanged. This behavior is explained by: i) the higher switching delays of most components of the video subsystem, which make switching to spares problematic even if the amount of time tolerated by the audience is higher, and ii) the higher number of components in the audio subsystem, which makes it appear less reliable in case switching is not possible (i.e., if the duration of "blackouts" tolerated by the audience is lower).

Finally, Figure 5i compares the effect of adding spares on the perceived reliability with respect to the different quality levels. Depending on the target QoE level and reliability requirements, different stage setups could be deployed. For example, adding a spare to each component would allow to reach 96% (perceived) reliability for a show requiring the higher quality level for both audio and video. However, the same (perceived) reliability could be reached without spares for shows that require the maximum quality for audio or video only. The results also show that, in the considered scenario, a reliability of 98% can be reached only if the performance is not video-based, i.e., if a degraded video quality can be accepted. Finally, a reliability greater than 99% can be achieved only if targeting a degraded quality level like Q_{22} or lower.

4.2.4 Comparison between "traditional" reliability and perceived reliability

In this section we provide an overview of the relationship between "traditional" reliability, and the *perceived reliability* metric defined in this paper. Following from its definition, perceived reliability is always equal to or greater than "traditional" reliability evaluated for the same setup. More in detail, the classical definition of reliability can be considered a particular instance of the perceived reliability metric, in which all the tolerable duration parameters ϑ are set to zero, i.e., failures of any duration are not tolerable at all.

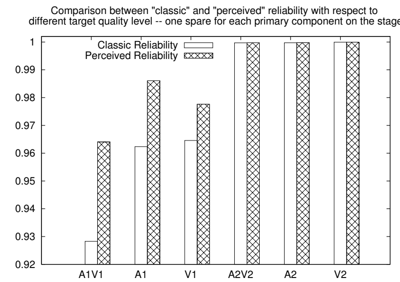


Figure 6: Comparison between $\mathcal{R}(t)$ and $\hat{\mathcal{R}}(t)$.

In Figure 6 traditional reliability ($\mathcal{R}(t)$) and perceived reliability ($\hat{\mathcal{R}}(t)$) are compared on the same setup, in which one spare is available for each primary component on the stage. Results confirm that, as expected, $\mathcal{R}(t)$ is always lower than $\hat{\mathcal{R}}(t)$, thus confirming the correctness of the modeling framework. By comparing results in Figure 6 with those in Figure 5i it can be seen that the value of traditional reliability with one spare for each component, and the perceived reliability without spares are roughly the same. This is caused switching delays being non-negligible: when using the traditional reliability metric the time required to switch to spare components is considered as interruption of service delivery.

4.3 Discussion

The obtained results demonstrate how the proposed framework can be used to provide useful recommendation to stage engineers and help in determining the dependability budget for such stage shows. Based on Figures 5a–5d and 5f we suggest adding spares to less reliable components only, and no more than two spares each. Figure 5g suggests the adoption of automatic switch over mechanism; Figure 5e suggests that using more reliable *wired* microphones does not improve the overall reliability (if spares are available), while it improves significantly if more reliable *wireless* microphones are employed. Based on Figure 5i the perceived reliability with respect to maximum audio and video quality cannot be improved over 0.965 by providing only hardware redundancy. Perhaps, information redundancy or/and software redundancy need to be introduced as well.

5. RELATED WORK

Recently emerging tele-immersive applications are designed to provide sophisticated features such as extensive configurability, high-resolution audio and video. These systems typically include a multitude of specialized hardware and software components. To meet the bandwidth requirements of the high resolution audio and video, these systems are designed over the advanced networking infrastructure. Accordingly, network packet losses have a smaller impact on the dependability of tele-immersive applications compared to malfunctioning of individual components.

[19], and [20] have shown that tele-immersive benefit the artist community to a great extent. [19] developed a distributed 3D-tele-immersive system, where the motion of a human body is captured and shared in a virtual space among the participants. The architecture is mainly designed to deal with video streams without considering audio or sensor streams. [20] presented the experimental results of performing collaborative dancing in the context of a 3D-tele-immersive system. The results conclude that a collaborative dance performance over network is feasible.

To the best of our knowledge, none of the existing tele-immersive applications considered failure scenarios in the design of their architecture. For this reason, the applications developed on top of these architectures do not provide maximum QoE to the end users in presence of failures.

Furthermore, the standards [2], and [1] established for evaluating the QoE of traditional audio/video teleconferencing applications do not pay significant attention to the faults of the client-side hardware components (such as microphone, workstations, etc.,) compared to the network level faults. Hence, they are not appropriate for evaluating the QoE in tele-immersive applications. A number of additional reasons for the non-applicability of the existing standards are discussed in [10].

The commonly used way to quantify the QoE in multimedia applications is to use a subjective assessment method [10]. In this method, a large number of people with various age differences are requested to rate the performance from 1 to 5 with 5 as the highest score for the best performance and mean of these scores are taken. This method is ineffective for developing reliable architecture to provide maximum QoE as it is expensive and consumes a lot of time to conduct the survey and performances. The other limitations of this method are proposed in [11].

[8] has quantified the QoE through a pentagram modeling framework. However, the modeling framework is limited to VoIP services and does not capture the complex characteristics of the tele-immersive applications. [6] proposed a pseudo-subjective quality assessment method to quantify the QoE using a neural network concept. The model only considers network failures.

In order to support early design decisions for providing maximum QoE, a modeling framework is required to compare the architectural variants. Our proposed perceived reliability concept, and the related modeling framework provides an effective method to *quantify* the QoE for users of tele-immersive applications in presence of client-side failures. While developed for the WO system, our approach can be exploited for the evaluation of QoE in any live distributed stage show with diverse audiences. More generally, the approach is directly applicable to a wide range of multimedia applications including e.g, video conferencing or online gaming, but also critical multimedia applications, e.g., distributed collaborative computer-assisted surgery [12].

6. CONCLUSION

Towards understanding the perceived QoE in presence of failures in distributed interactive multimedia applications, we proposed a new approach based on the "perceived reliability", which takes into account human perception limits and subjective satisfaction. We then designed and implemented a modeling framework using Stochastic Activity Networks to evaluate such metrics and demonstrate the feasibility of our approach. The obtained results have provided useful insights to aid technicians in the deployment of World Opera performances.

This work can be used as basis for a thorough evaluation of QoE in WO performances. In future, we would like to extend the framework to consider propagation and interdependencies between failures occurring at different stages, and evaluate the QoE perceived by different kind of users, including various types of performing artists (such as dancer, singer, musician, conductor, and story narrator), or audience located in different parts of the stage. Finally, we aim to extend this approach to other multimedia applications, including critical and safety-critical applications.

7. REFERENCES

[1] ITU-G.1070. Opinion model for video-telephony applications, 2007.

[2] ITU-G.107. The E-model, a computational model for use in transmission planning, 2008.

[3] The World Opera. `http://theworldopera.org`, September 2012.

[4] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr. Basic concepts and taxonomy of dependable and secure computing. *IEEE Trans. Depend. Secur. Comput.*, pages 11–33, 2004.

[5] G. Ciardo, R. German, and C. Lindemann. A characterization of the stochastic process underlying a stochastic petri net. In *5th International Workshop on Petri Nets and Performance Models*, pages 170–179, 1993.

[6] A. P. C. da Silva, M. Varela, E. de Souza e Silva, R. M. M. Leão, and G. Rubino. Quality assessment of interactive voice applications. *Computer Networks*, 52(6):1179–1192, 2008.

[7] S. Gaonkar, K. Keefe, R. Lamprecht, E. Rozier, P. Kemper, and W. H. Sanders. Performance and dependability modeling with Möbius. *SIGMETRICS Perform. Eval. Rev.*, 36:16–21, March 2009.

[8] Y. Gong, F. Yang, L. Huang, and S. Su. Model-based approach to measuring quality of experience. In *Proc. of the 1st Int. Conference on Emerging Network Intelligence*, EMERGING '09, pages 29–32, 2009.

[9] T. Hampel, T. Bopp, and R. Hinn. A peer-to-peer architecture for massive multiplayer online games. In *Proc. ACM NetGames '06*, 2006.

[10] Z. Huang, A. Arefin, P. Agarwal, K. Nahrstedt, and W. Wu. Towards the understanding of human perceptual quality in tele-immersive shared activity. In *Proceedings of the 3rd Multimedia Systems Conference*, MMSys '12, 2012.

[11] H. Knoche, H. G. D. Meer, and D. Kirsh. Utility curves: Mean opinion scores considered biased. In *Proc. of 7th Int. Workshop on Quality of Service*, pages 12–14, 1999.

[12] H. Lufei, W. Shi, and V. Chaudhary. M-CASEngine: a collaborative environment for computer-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*, 1 SUPP/1:447–448, 2006.

[13] J. Meyer. On evaluating the performability of degradable computing systems. *IEEE Transactions on Computers*, C-29(8):720 –731, aug. 1980.

[14] N. Raghavan, L. Montecchi, N. Nostro, A. Bondavalli, R. Vitenberg, and H. Meling. Modeling QoE in distributed multimedia applications. Technical report, September 2012.

[15] W. Sanders and J. Meyer. Reduced base model construction methods for stochastic activity networks. *IEEE Journal on Selected Areas in Communications*, 9(1):25–36, jan 1991.

[16] W. H. Sanders and J. F. Meyer. Stochastic activity networks: formal definitions and concepts. In *Lectures on formal methods and performance analysis*, pages 315–343. 2002.

[17] K. S. Trivedi. *Probability and Statistics with Reliability, Queueing, and Computer Science Applications, 2nd Edition*. Wiley-Interscience, Oct. 2001.

[18] W. Wu, A. Arefin, R. Rivas, K. Nahrstedt, R. Sheppard, and Z. Yang. Quality of experience in distributed interactive multimedia environments: toward a theoretical framework. In *Proc. ACM Multimedia'09*, pages 481–490, 2009.

[19] W. Wu, Z. Yang, D. Jin, and K. Nahrstedt. Implementing a distributed tele-immersive system. In *Proc. of the 10th IEEE ISM'08*, pages 477–484.

[20] Z. Yang, B. Yu, W. Wu, R. Diankov, and R. Bajscy. Collaborative dancing in tele-immersive environment. In *ACM Multimedia*, pages 723–726, 2006.