

# Quantifying Computational Reliability of Task Assignment in Extreme Edge Computing

MHD Saria Allahham, *Member, IEEE* and Hossam S. Hassanein, *Fellow, IEEE*

**Abstract**—This paper presents a reliability analysis framework for distributed computing in extreme edge computing (XEC) with limited information availability. XEC pushes computation to the outermost boundaries of networks by leveraging consumer-owned devices, known as Extreme Edge Devices (XEDs). Unlike traditional distributed systems with defined computational resource states, XEC operates under uncertainty due to consumer device usage patterns, varying computational capacities, and local scheduling algorithms. In this work, we address discrete task assignment particularly. The framework analyzes scenarios for computational reliability assessments with minimal knowledge of XED capabilities and service requirements. The framework adapts to different levels of available information, from operational limits to historical performance data, providing refined reliability estimates. The aim of this work is to provide generalized reliability models for distributed computing in XEC that allow decision-makers (e.g. service orchestrators) to make informed decisions about task allocation, service placement, and resource allocation under uncertainty in XEC. Simulations and experimental analysis demonstrate the framework's effectiveness in estimating reliability under various system conditions.

**Index Terms**—reliability modeling, extreme edge computing, task assignment

## I. INTRODUCTION

The rapid proliferation of Artificial Intelligence (AI) has spurred a demand for real-time, low-latency inferencing to power a new generation of intelligent applications, from autonomous systems to interactive services [1]. The process of AI inference, where a trained model makes predictions on new data, is the operational cornerstone that delivers the value of AI to end-users. However, deploying these models, particularly large-scale models like Large Language Models (LLMs), presents significant computational challenges. These models are often compute and memory intensive, and executing them with the minimal latency required for real-time interaction strains traditional centralized cloud architectures [2], [3]. One emerging paradigm to address these compute-intensive workloads is the Computing Power Network (CPN) [4]. A CPN envisions a future where computational resources are treated as a tradable utility, seamlessly connecting heterogeneous and ubiquitous computing power from the cloud to the edge to meet diverse application demands. This model is particularly well-suited for a future of pervasive AI, as it provides a framework for orchestrating vast, distributed resources to handle intensive tasks. The primary enabler for realizing such a vision is Extreme Edge Computing (XEC), which extends the computing continuum to the vast and often

untapped resources of consumer-owned devices [5], [6], [7]. XEC harnesses the collective power of Extreme Edge Devices (XEDs) such as smartphones, wearables, and IoT sensors to bring computation to the absolute network periphery, closest to the data source [8], [9]. However, the foundational challenge of building a dependable system like a CPN upon an XEC substrate is the inherent unreliability of the resources. Unlike enterprise-owned edge servers, XEDs are characterized by profound uncertainty; their computational availability is subject to user behavior, mobility, resource constraints like battery life, and local, uncoordinated scheduling [10]. This confluence of high, dynamic demand from AI applications with the inherent unpredictability and intermittent availability of XEDs stresses the system's ability to provide consistent performance and meet strict latency requirements. Consequently, for the XEC paradigm to be viable, especially for demanding AI services, the dependability of its constituent computational resources must be quantifiable.

The requirement to quantify the dependability of individual XEDs brings the concept of computational reliability to the forefront. Assessing this type of reliability in the XEC environment exposes a critical gap in existing literature. Previous works often focus on connectivity or hardware failures and typically assume deterministic knowledge of a device's computational state [11]. These assumptions are invalid in XEC, where the stochastic nature of available computational resources, unpredictable user behavior patterns, and limited system state information create unique challenges that existing reliability models cannot address. Our work addresses this crucial gap by developing a framework for quantifying computational reliability under uncertainty, specifically tailored for XEDs. We define computational reliability as the probability of successfully completing a task within its deadline, deliberately isolating this metric from other factors like network connectivity.

Our work addresses this crucial gap by developing a framework for modeling **computational reliability** under uncertainty, specifically tailored for systems where computational resources are provided by consumer-owned devices. Addressing this is not an incremental improvement but a necessity; without a reliable method to quantify computational performance, the entire XEC paradigm risks failure, as it cannot guarantee the service quality required for the data-intensive and latency-sensitive applications it is meant to enable. Our focus is on task assignment scenarios where computation requests are discrete and time-bounded. While this scenario shares similarities with traditional distributed computing, the inherent characteristics of XEDs heterogeneity, dynamic resource availability, and limited state information

MHD Saria Allahham and Hossam Hassanein are with School of Computing, Queen's University, ON, Canada. (emails: 20msa7@queensu.ca and hossam@cs.queensu.ca)

introduce complexity.

We particularly emphasize modeling single worker reliability purely from a **computational perspective**, that is, we take the computational reliability to be probability of successfully completing a task within its deadline, deliberately isolating it from other factors like network connectivity. By providing a quantitative model for this specific uncertainty, it equips decision-makers, such as service orchestrators, with the tools to make informed, risk-aware decisions for instance, by comparing the calculated reliability of several available workers to select the most suitable one for a critical task, thereby optimizing task allocation and resource management under uncertainty. The main contributions of this work are as follows:

- We establish a foundational framework for quantifying computational reliability for task assignment in XEC systems operating under uncertainty.
- We derive mathematically tractable, closed-form, and approximate closed-form expressions for computational reliability as a function of the task time-deadline and task requirements.
- We introduce a data-driven refinement process, and proving that incorporating historical observations via Maximum Likelihood Estimation (MLE) reduces reliability estimation uncertainty.

The remainder of this paper is organized as follows: In Section II, we discuss the previous works. Section III presents the system model. Section IV details the modeling of task assignment reliability, including approaches for both with minimal information and with historical data scenarios. Section V presents simulation and experimental results validating our approach. Finally, we conclude in Section VII.

## II. RELATED WORK

The evolution of edge computing research has witnessed significant developments in addressing the challenges of moving computation closer to data sources. The work by Shi et al. [2], [12] established the initial vision and challenges of edge computing, while subsequent surveys [3] comprehensively mapped the landscape of mobile edge computing from architectural and communication perspectives. As edge computing evolved, researchers identified limitations in traditional edge server approaches, particularly for ultra-low latency applications [13]. The democratization of extreme edge resources has been extensively studied, with seminal works by Tourani et al. [14] and Portilla et al. [5] laying the foundations, establishing the fundamental architecture and implementation requirements of XEC systems. These works were complemented by the authors in [8], who developed adaptive algorithms specifically for wearable devices in XEC environments, focusing on power usage optimization and processing efficiency for pervasive computing scenarios. Resource allocation in XEC has emerged as a critical research focus, with various approaches addressing different aspects of the challenge. In [15], the authors introduced the Community-Oriented Resource Allocation (CORA) scheme, which uniquely combines resource optimization with privacy preservation through community-based constraints. This work was complemented by [16],

which proposed the Optimal Proactive Resource Allocation (OPRA) framework, introducing proactive resource management to minimize system delays in XEC. The fairness aspect of resource allocation was addressed by the authors in [17] through their Multitiered Worker-Oriented Resource Allocation (MWORA) scheme, while others in [18] focused on cost-efficient recruitment through their Price-based Compute Clusters Recruitment (PCCR) approach. Moreover, the dynamic nature of XEC resources has prompted significant research into resource characterization and prediction. The authors in [19] developed the Usage-based Worker Resource Characterization (U-WORC) scheme, employing clustering techniques for analyzing resource usage patterns. This work was extended by Kain et al. [20], [21] through the Resource Usage Multi-Step Prediction (RUMP) scheme, which leverages the Hierarchical Dirichlet Process-Hidden Semi-Markov Model for improved forecasting accuracy. While these studies address resource availability and allocation, they do not directly model the inherent computational reliability of the individual workers providing these resources.

The concept of reliability in edge computing, particularly in XEC systems, has been explored through various lenses in recent research such as: connectivity [22], [23], security [24], [25], and most relevantly, computation [26], [27]. For instance, in [26], reliability is associated with offloading computationally intensive components from mobile devices to edge servers, addressing challenges such as limited power and server capacity constraints. This study focuses on maximizing computing tasks while minimizing energy consumption, and developing strategies to ensure minimum latency in tasks allocated among multiple devices and servers. In [28], the authors discuss reliability in the context of ultra-reliable, low-latency edge computing services. This paper emphasizes the need for distributed edge decision-making services to adapt swiftly to network dynamics. The connectivity reliability was studied for mission-critical applications like virtual reality, vehicle-to-everything communication, and edge AI, where any delay or failure can have significant consequences. In [29], the reliability of a computer network, especially in a distributed network comprising IoT devices, edge, and cloud servers, is examined. The focus is on the successful processing and transmission of data within this network, considering the multi-state nature of transmission lines and the diverse capacities of network components. The work in [27] brings reliability into the field of the Internet of Vehicles, stressing the importance of ensuring high-reliability levels for latency-sensitive applications. Here, reliability is considered in terms of both processing nodes and communication links, introducing mechanisms like task allocation and reprocessing to maintain service reliability. Whereas [30] explores reliability in large-scale edge computing systems, particularly in the context of service composition. The study addresses the challenges posed by failures, recoveries, and complex architecture by proposing simulation-based optimization for reliability-aware service composition. In [31], the authors tie reliability in terms of failure to the bandwidth consumption of IoT applications in edge computing environments. The study investigates the equilibrium between maximizing reliability and minimizing

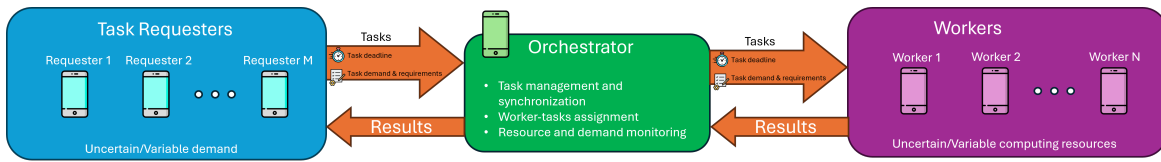


Fig. 1. XEC system model

bandwidth consumption during task offloading.

Reliability modeling has also been studied extensively in volunteer computing and grid systems. In volunteer computing platforms such as BOINC, reliability models primarily address device availability (i.e., whether a device remains connected long enough to complete a task) and result correctness through redundant computation and majority voting [32], [33]. Similarly, grid computing reliability research has focused on failure rates, checkpoint strategies, and task replication under hardware faults [34]. Heavy-tailed execution times in heterogeneous distributed systems have been reported in [35], motivating the use of flexible distributional models. While these works share surface-level similarities with our problem, they differ from the XEC setting in important ways. First, volunteer and grid systems typically operate with accumulated task completion statistics and centralized scheduling histories, whereas XEC must handle scenarios where such information is unavailable or minimal due to devices joining and leaving dynamically. Second, the reliability focus in volunteer computing is on device availability and result correctness, whereas our work focuses on computational performance reliability: given that a device is available, what is the probability it completes the task within the deadline under uncertain capacity? Third, our framework provides a unified progression from maximum-entropy baselines to data-driven estimates as information accumulates, addressing the unique lifecycle of participants in XEC.

Beyond volunteer computing, deadline-constrained task scheduling in distributed systems has been widely studied. The authors in [36] proposed cost-driven scheduling algorithms for deadline-constrained workflows on utility grids, while [37] addressed deadline-constrained workflow scheduling in IaaS cloud environments. However, these works assume known or deterministic computational capacities, which does not hold in XEC. Additionally, mobile device resource variability has been characterized in measurement studies such as [38], which reported significant fluctuations in mobile device performance due to thermal throttling, background processes, and battery states. Our framework directly addresses this variability by modeling capacity as a random variable and providing reliability estimates under such uncertainty.

### III. SYSTEM MODEL

We consider an Extreme Edge Computing (XEC) system as in Figure 1, which comprises three main components: service providers/workers, service requesters, and an orchestrator. The service providers, also referred to as workers, offer their computational resources to perform tasks offloaded by the orchestrator. We consider that the workers can be consumer-owned

devices such as mobile phones or wearables. As such, workers have unpredictable computational behavior, such that their computational resources change randomly. Service requesters are entities that generate computational tasks and offload them to the orchestrator. These requesters can be thought of as users who require computational resources based on their specific needs. They may include individuals using mobile phones, IoT sensors collecting data, or applications running on various platforms. The orchestrator serves as the central coordination entity in the XEC system. Its primary functions include synchronization, task collection and assignment, and worker resources and requesters task demand monitoring. To fulfill these roles, the orchestrator continuously collects and analyzes data from both workers and requesters. For workers, it tracks their available computational capacity, and current workload. For requesters, it monitors incoming task requests, their computational requirements, and any associated deadlines or priorities.

In an XEC system, we focus our work on task assignment, which refers to the process of delegating a discrete, one-time computational task from a requester to a worker with a specified deadline. These tasks are typically finite in duration and have clear start and end points. The reliability of task assignment is primarily concerned with the successful completion of the task within the given time constraint. For ease of reference, table I summarizes the key notation used throughout this paper.

TABLE I  
SUMMARY OF KEY NOTATION

Symbol	Description
$T$	Random variable for task completion time
$T_d$	Task time deadline
$R(T_d)$	Reliability function
$C$	Computational capacity of a worker (RV)
$D$	Computational demand of a task (RV)
$C_{\min}, C_{\max}$	Min/max computational capacity bounds
$D_{\min}, D_{\max}$	Min/max task demand bounds
$C_{\text{range}}, D_{\text{range}}$	$C_{\max} - C_{\min}, D_{\max} - D_{\min}$
$\alpha$	Scale parameter (expected $C/D$ ratio)
$\xi$	Shape parameter of the GPD
$\mu_C, \sigma_C$	Mean and std. dev. of capacity (historical)
$\mu_D, \sigma_D$	Mean and std. dev. of demand (historical)
$C_{\min}, C_{\max}$	Standardized capacity bounds
$X_{\min}, X_{\max}$	Standardized truncation bounds
$\phi(\cdot), \Phi(\cdot)$	Standard normal PDF and CDF
$H(\cdot)$	Entropy

### IV. MODELING OF TASK ASSIGNMENT RELIABILITY

In this section, we focus on formalizing the concept of reliability of a worker for task assignment given a specific time

deadline, and explore different modeling approaches based on the information available at the orchestrator.

The reliability of task assignment can be thought of as the probability that a given task will be completed within its designated deadline. Mathematically, we model the reliability function  $R(T_d)$  of a worker as:

$$R(T_d) := F_T(T_d) = P(T \leq T_d) = \int_0^{T_d} f_T(x) dx \quad (1)$$

where  $T$  is a random variable (RV) representing the actual task completion time,  $T_d$  is the specified deadline for task completion, and  $f_T(x)$  is the probability density function of the task completion time.

While several works in the literature have considered exponential distributions for task completion times, recent studies suggest that computational task times in many systems follow heavier-tailed distributions [35]. However, considering the heterogeneous nature of XEDs and the wide variety of tasks at the edge, each with different demands and characteristics, we opt for a more generalized model, which is the Generalized Pareto Distribution (GPD) [39]. The GPD provides greater flexibility in modeling as it encompasses both exponential and all types of Pareto distributions. Notably, when  $\xi = 0$ , the GPD reduces to the exponential distribution commonly assumed in prior work, while  $\xi > 0$  captures the heavy-tailed (Pareto-type) behavior reported in [35]. This single-parameterization flexibility makes the GPD particularly suitable for the diverse workloads encountered in XEC, as opposed to fixed distributional forms such as log-normal or mixture models. The cumulative distribution function of the GPD is given by:

$$F(T; \xi, \alpha) = \begin{cases} 1 - (1 + \xi\alpha T)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1 - e^{-\alpha T} & \text{if } \xi = 0 \end{cases} \quad (2)$$

Here,  $\alpha > 0$  is the scale parameter, indicating the spread of the distribution, while  $\xi$  is the shape parameter determining the tail behavior.

The GPD's versatility allows us to model a wide range of scenarios, from light-tailed to heavy-tailed distributions, making it particularly suitable for the heterogeneous and often unpredictable computational behavior of XEDs. The shape parameter  $\xi$  determines the asymptotic tail of the distribution, while the scale parameter  $\alpha$  indicates how quickly the reliability increases over time, or the rate at which the reliability approaches the asymptotic tail. Consequently, this approach provides a more realistic representation of reliability in XEC systems compared to simpler models like exponential or single-parameter distributions.

It is worth noting that the  $k^{\text{th}}$  statistical moment of the GPD is only defined for  $\xi < 1/k$ . For instance, the first moment, or expected value, of the GPD is given by:

$$\mathbb{E}[T] = \begin{cases} \frac{1}{\alpha} & \xi = 0 \\ \frac{1}{\alpha(1-\xi)} & 0 < \xi < 1 \\ \text{undefined} & \xi \geq 1 \end{cases} \quad (3)$$

When  $\xi \geq \frac{1}{k}$ , we can still study the reliability of an XED, even though estimating the mean task execution time becomes

challenging due to the heavy tail of the distribution. This aligns with real-world scenarios where some tasks may take an exceptionally long time to execute or may not complete at all, making the XED less reliable.

In our work, we focus on modeling the parameter  $\alpha$ , which plays a crucial role in characterizing the reliability. This parameter represents the rate at which tasks are executed and directly influences how quickly reliability increases over time. Specifically,  $\alpha$  can be expressed as the expected ratio between an XED's computational capacity and the task's computational demand:

$$\alpha = \mathbb{E} \left[ \frac{C}{D} \right] \quad (4)$$

where  $C$  represents the computational capacity (i.e., computational resources) of an XED and  $D$  denotes the computational task demand. Both  $C$  and  $D$  are inherently non-negative and independent RVs. The independence assumption is justified by the fact that  $C$  and  $D$  originate from fundamentally separate entities:  $C$  is determined by the worker's device state (influenced by local processes, battery level, and thermal conditions), while  $D$  is determined by the requester's application requirements (input data size, algorithm complexity, and client-specific parameters). Since a worker's current resource state does not influence what computation a requester needs, and vice versa, independence is a reasonable and standard assumption in the task scheduling literature. The parameter  $\alpha$  thus serves as a critical performance indicator, encapsulating the fundamental relationship between resource availability and workload requirements. A higher value of  $\alpha$  signifies a greater expected surplus of computational capacity relative to the task's demand, directly translating to a higher probability of timely task completion. It distills the interplay between a worker's state and a task's needs into a single metric for reliability assessment.

Modeling the capacity as an RV reflects the dynamic nature of XEDs, which may experience fluctuations in available resources. These fluctuations can arise from various factors, including local concurrent processes competing for resources, low battery levels, or thermal constraints that necessitate the Core Processing Unit (CPU) throttling. Whereas the task demand is modeled as an RV to account for the diversity and unpredictability of edge computing workloads. The tasks can vary significantly in their computational requirements, depending on factors such as input data size, algorithm complexity, or client-specific parameters.

We explore two distinct scenarios for modeling task assignment reliability: one with minimal information and another with historical data. These two scenarios were deliberately chosen as they represent the fundamental poles on a spectrum of information availability that an orchestrator might face. The minimal information model establishes a crucial baseline for reliability when a worker or task is new to the system, relying only on declared operational bounds (e.g., min/max capacity). This represents the highest degree of uncertainty. Conversely, the historical data model represents a more evolved and informed state, where past interactions provide empirical evidence to refine reliability estimates and reduce uncertainty.

By defining these two extremes, our framework offers a generalized and adaptable approach; any participant can be initialized with the minimal model and seamlessly transition to the historical data model as interactions accumulate, making the framework broadly applicable to the dynamic nature of XEC.

### A. Modeling with Minimal Information

In this scenario, we address the challenge of estimating the reliability metric with limited information. To participate in the XEC system, each entity must provide a minimum set of parameters to the orchestrator. This set, termed Minimal Information (MI), is essential for establishing an initial computational reliability estimation. For workers, the MI comprises the bounds of their allocable computational resources, specifically the maximum and minimum computational capacities. Conversely, requesters must provide the upper and lower limits of their computational demands. The MI, which can be conveniently provided by both workers and requesters, serves as a foundation for deriving initial reliability estimates when historical execution information is unavailable.

As such, we model both the computational capacity  $C$  and task demand  $D$  as uniformly distributed RVs. The selection of the uniform distribution represents the worst case scenario regarding a new participant without any prior history, as it yields the highest possible uncertainty (maximum entropy) [40]. When the only available information consists of bounds (minimum and maximum values), no prior information exists that would allow treating the uncertainty between  $C_{\min}$  and  $C_{\max}$  differently, necessitating adherence to the principle of indifference [41]. The uniform distribution's application in this context is particularly compelling as it requires only the knowledge of lower and upper bounds, aligning with our MI framework where workers and requesters provide only these boundary values. Moreover, any alternative distribution choice would impose unjustifiable assumptions about the probability density within these bounds, which cannot be supported given the limited information available.

Let  $C \sim U(C_{\min}, C_{\max})$  and  $D \sim U(D_{\min}, D_{\max})$ , where  $C_{\min}$  and  $C_{\max}$  are the minimum and maximum computational capacities, and  $D_{\min}$  and  $D_{\max}$  are the minimum and maximum task demands, respectively. For ease of analysis, we define the following:

$$C_{\text{range}} = C_{\max} - C_{\min}, \quad D_{\text{range}} = D_{\max} - D_{\min} \quad (5)$$

To derive the reliability metric, we need to calculate the mean and variance of the ratio  $C/D$ .

According to Melvin [42], for non-negative, independent random variables  $X$  and  $Y$ , the expected value of their ratio and their squared ratio is given by, respectively:

$$\mathbb{E} \left[ \frac{X}{Y} \right] = \mathbb{E}[X] \mathbb{E} \left[ \frac{1}{Y} \right], \quad \mathbb{E} \left[ \left( \frac{X}{Y} \right)^2 \right] = \mathbb{E}[X^2] \mathbb{E} \left[ \frac{1}{Y^2} \right] \quad (6)$$

Accordingly, we have:

$$\alpha = \mathbb{E} \left[ \frac{C}{D} \right] = \mathbb{E}[C] \mathbb{E} \left[ \frac{1}{D} \right], \quad (7)$$

and for a uniform RV  $X \sim U(a, b)$ , we know that:

$$\mathbb{E}[X] = \frac{a+b}{2}, \quad \mathbb{E} \left[ \frac{1}{X} \right] = \frac{\log(b) - \log(a)}{b-a} \quad (8)$$

Substituting these into our equation for  $\alpha$ , we derive:

$$\alpha_{\text{MI}} = \mathbb{E} \left[ \frac{C}{D} \right] = \frac{(C_{\max} + C_{\min}) \log \left( \frac{D_{\max}}{D_{\min}} \right)}{2D_{\text{range}}} \quad (9)$$

To fully characterize the distribution of  $C/D$ , we also need to calculate its variance. We can do this by first calculating  $\mathbb{E}[(C/D)^2]$  and then using the relation:

$$\sigma_{\text{MI}}^2 = \mathbb{E} \left[ \left( \frac{C}{D} \right)^2 \right] - \left( \mathbb{E} \left[ \frac{C}{D} \right] \right)^2 \quad (10)$$

To derive the expression for  $\mathbb{E} \left[ \left( \frac{C}{D} \right)^2 \right]$ , we follow a similar approach as before, leveraging the independence of  $C$  and  $D$ . From (5), we have:

$$\mathbb{E} \left[ \left( \frac{C}{D} \right)^2 \right] = \mathbb{E}[C^2] \mathbb{E} \left[ \frac{1}{D^2} \right] \quad (11)$$

and for a uniform RV  $X \sim U(a, b)$ , we know that:

$$\mathbb{E}[X^2] = \frac{b^3 - a^3}{3(b-a)}, \quad \mathbb{E} \left[ \frac{1}{X^2} \right] = \frac{1}{b-a} \left( \frac{1}{a} - \frac{1}{b} \right) \quad (12)$$

Substituting these into our equation, we get:

$$\mathbb{E} \left[ \left( \frac{C}{D} \right)^2 \right] = \frac{(C_{\max}^3 - C_{\min}^3) \left( \frac{1}{D_{\min}} - \frac{1}{D_{\max}} \right)}{3C_{\text{range}}D_{\text{range}}} \quad (13)$$

The derivation of both the mean  $\alpha_{\text{MI}}$  and standard deviation  $\sigma_{\text{MI}}$  is necessary because these parameters fully characterize the distribution of  $C/D$  under our assumptions. This information is crucial for estimating the reliability of task assignment, as it provides a measure of both the expected performance (through the mean) and the variability in this performance (through the standard deviation).

By leveraging these derived equations, with  $T_d$  being the task time deadline, we can now derive the reliability metric for a worker with MI by substituting Eq. (9) in Eq. (2) as follows:

$$R_{\text{MI}}(T_d) = 1 - \left( 1 + \frac{\xi(C_{\max} + C_{\min}) \log \left( \frac{D_{\max}}{D_{\min}} \right) T_d}{2D_{\text{range}}} \right)^{-\frac{1}{\xi}} \quad (14)$$

for the case where  $\xi \neq 0$ , and:

$$R_{\text{MI}}(T_d) = 1 - \exp \left( - \frac{(C_{\max} + C_{\min}) \log \left( \frac{D_{\max}}{D_{\min}} \right) T_d}{2D_{\text{range}}} \right) \quad (15)$$

for the case where  $\xi = 0$ , where  $\exp(\cdot)$  is the exponential function.

## B. Modeling with Historical Data

In this scenario, we consider cases where the orchestrator has accumulated interaction data over time or the worker has a record of interactions. Through repeated task assignment interactions, the orchestrator builds a historical record of computational capacities and task demands. This historical data reveals patterns in how workers perform tasks and how requesters demand computational resources. The motivation for estimating parameters from historical data stems from fundamental statistical principles. While the MI approach provides a conservative baseline using maximum entropy principles, historical data allows us to leverage the law of large numbers and central limit theorem to obtain more precise estimates of the worker behavior. As the number of observations increases, sample statistics converge to their true population parameters, enabling more accurate reliability predictions. This transition from maximum entropy (uniform distribution) to data-driven estimation aligns with Bayesian updating principles, where initial uninformative priors are refined through observed evidence.

We model the observations using truncated normal distributions, bounded by the MI-specified minimum and maximum values. This choice is motivated by several theoretical considerations: (1) The central limit theorem suggests that the sum of many independent random effects tends toward normality (2) The truncation naturally incorporates the physical constraints (i.e., the minimum and maximum capacity) while preserving the theoretical properties of normal distributions, and (3) It provides analytical tractability for parameter estimation while capturing both symmetric and moderately skewed distributions common in computational systems. The observations are treated as independent and identically distributed (i.i.d.) in our model. The independence assumption is justified through the temporal separation between task executions, while the identical distribution assumption holds for measurement windows where worker device characteristics and workload patterns remain relatively stable. We note that stationarity is assumed within a measurement window, not globally. In practice, the MLE parameters can be periodically re-estimated with fresh observations, creating a sliding-window approach that gracefully accommodates non-stationarity over longer time horizons.

Given a set of i.i.d. observations  $x_1, x_2, \dots, x_n$  representing the history of capacity or task demand, assumed to be drawn from a truncated normal distribution with unknown parameters  $\mu$  and  $\sigma$ , and known truncation points  $X_{\min}$  and  $X_{\max}$ , the Maximum Likelihood Estimates (MLE) of  $\mu$  and  $\sigma$  can be obtained as follows:

$$\mu_{\text{MLE}}, \sigma_{\text{MLE}} = \arg \max_{\mu, \sigma} \log \mathcal{L}(\mu, \sigma; X_{\min}, X_{\max} | x_1, \dots, x_n) \quad (16)$$

where

$$\begin{aligned} \log \mathcal{L}(\mu, \sigma; X_{\min}, X_{\max} | x_1, \dots, x_n) = & \\ & -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \\ & - n \log [\Phi(\bar{X}_{\max}) - \Phi(\bar{X}_{\min})], \end{aligned} \quad (17)$$

and  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal distribution, subject to the constraints  $\sigma > 0$  and  $X_{\min} \leq \mu \leq X_{\max}$ .

Let  $C \sim \text{TNORM}(\mu_C, \sigma_C, C_{\min}, C_{\max})$  and  $D \sim \text{TNORM}(\mu_D, \sigma_D, D_{\min}, D_{\max})$ , where  $\mu_C, \sigma_C$  and  $\mu_D, \sigma_D$  are the mean and standard deviation of the computational capacity and task demand, respectively.

For a truncated normal distribution  $X \sim \text{TNORM}(\mu, \sigma, X_{\min}, X_{\max})$ , the first moment, the second moment, and variance are given by, respectively:

$$\mathbb{E}[X] = \mu + \sigma \frac{\phi(\bar{X}_{\min}) - \phi(\bar{X}_{\max})}{\Phi(\bar{X}_{\max}) - \Phi(\bar{X}_{\min})}, \quad (18)$$

$$\mathbb{E}[X^2] = \text{Var}(X) + \left( \mu + \sigma \frac{\phi(\bar{X}_{\min}) - \phi(\bar{X}_{\max})}{\Phi(\bar{X}_{\max}) - \Phi(\bar{X}_{\min})} \right)^2, \quad (19)$$

$$\begin{aligned} \text{Var}(X) = \sigma^2 \left( 1 + \frac{\bar{X}_{\min} \phi(\bar{X}_{\min}) - \bar{X}_{\max} \phi(\bar{X}_{\max})}{\Phi(\bar{X}_{\max}) - \Phi(\bar{X}_{\min})} \right. \\ \left. - \left( \frac{\phi(\bar{X}_{\min}) - \phi(\bar{X}_{\max})}{\Phi(\bar{X}_{\max}) - \Phi(\bar{X}_{\min})} \right)^2 \right), \end{aligned} \quad (20)$$

where  $\bar{X}_{\min} = \frac{X_{\min} - \mu}{\sigma}$ ,  $\bar{X}_{\max} = \frac{X_{\max} - \mu}{\sigma}$ ,  $\phi(z)$  and  $\Phi(z)$  are the probability density function and cumulative distribution function of the standard normal distribution, respectively.

To derive the reliability metric, we need to calculate  $\mathbb{E}[C/D]$  and  $\mathbb{E}[(C/D)^2]$ . However, the evaluation of  $\mathbb{E}[1/X]$  for a truncated normal distribution is analytically intractable. Therefore, we propose to approximate it using a second-order Taylor series expansion.

For a random variable  $X$  with finite mean  $\mu$  and variance  $\sigma^2$ , the expectation of  $1/X$  can be approximated as:

$$\mathbb{E} \left[ \frac{1}{X} \right] \approx \frac{1}{\mu} + \frac{\text{Var}(X)}{\mu^3} + \epsilon_1 \quad (21)$$

where  $\epsilon_1 = O\left(\frac{1}{6} \frac{\mathbb{E}[(X-\mu)^3]}{\mu^4}\right)$  is the error term.

Let  $g(X) = 1/X$ . The second-order Taylor expansion of  $g(X)$  around  $\mu$  is:

$$g(X) \approx g(\mu) + g'(\mu)(X - \mu) + \frac{1}{2}g''(\mu)(X - \mu)^2 \quad (22)$$

Taking the expectation of both sides and noting that  $\mathbb{E}[X - \mu] = 0$ , we get:

$$\mathbb{E}[g(X)] \approx g(\mu) + \frac{1}{2}g''(\mu)\mathbb{E}[(X - \mu)^2] \quad (23)$$

Substituting  $g(\mu) = 1/\mu$ ,  $g''(\mu) = 2/\mu^3$ , and  $\mathbb{E}[(X - \mu)^2] = \text{Var}(X)$ , we arrive at the stated approximation.

Similarly, for a random variable  $X$  with finite mean  $\mu$  and variance  $\sigma^2$ , the expectation of  $1/X^2$  can be approximated as:

$$\mathbb{E} \left[ \frac{1}{X^2} \right] \approx \frac{1}{\mu^2} + \frac{3\text{Var}(X)}{\mu^4} + \epsilon_2 \quad (24)$$

where  $\epsilon_2 = O\left(\frac{1}{4} \frac{\mathbb{E}[(X-\mu)^4]}{\mu^6}\right)$  is the error term.

These derivations help us in deriving tractable approximations for the ratio of computational capacity to task demand. They provide second-order Taylor series approximations for the inverse moments that appear in our reliability calculations, allowing us to express  $\mathbb{E}[C/D]$  and  $\mathbb{E}[(C/D)^2]$  in terms of the means and variances of the truncated normal distributions. This is because these expectations are otherwise analytically intractable for truncated normal distributions. By leveraging these approximations and the independence between  $C$  and  $D$ , we can derive closed-form expressions for the key parameters needed to characterize the reliability metric.

Given  $C \sim \text{TNORM}(\mu_C, \sigma_C, C_{\min}, C_{\max})$  and  $D \sim \text{TNORM}(\mu_D, \sigma_D, D_{\min}, D_{\max})$ , the expectation  $\mathbb{E}[C/D]$ , i.e., the mean execution rate with historical data  $\alpha_{\mathbf{H}}$ , and  $\mathbb{E}[(C/D)^2]$  can be approximated as:

$$\alpha_{\mathbf{H}} = \mathbb{E}\left[\frac{C}{D}\right] = \mathbb{E}[C] \mathbb{E}\left[\frac{1}{D}\right] \approx \left(\mu_C + \sigma_C \frac{\phi(\bar{C}_{\min}) - \phi(\bar{C}_{\max})}{\Phi(\bar{C}_{\max}) - \Phi(\bar{C}_{\min})}\right) \left(\frac{1}{\mu_D} + \frac{\text{Var}(D)}{\mu_D^3}\right) \quad (25)$$

$$\mathbb{E}\left[\left(\frac{C}{D}\right)^2\right] = \mathbb{E}[C^2] \mathbb{E}\left[\frac{1}{D^2}\right] \approx \left(\text{Var}(C) + \left(\mu_C + \sigma_C \frac{\phi(\bar{C}_{\min}) - \phi(\bar{C}_{\max})}{\Phi(\bar{C}_{\max}) - \Phi(\bar{C}_{\min})}\right)^2\right) \left(\frac{1}{\mu_D^2} + \frac{3\text{Var}(D)}{\mu_D^4}\right) \quad (26)$$

where  $\bar{C}_{\min} = (C_{\min} - \mu_C)/\sigma_C$  and  $\bar{C}_{\max} = (C_{\max} - \mu_C)/\sigma_C$ . Consequently, we can estimate the XED reliability with a time deadline  $T_d$  as follows:

$$R_{\mathbf{H}}(T_d) \approx 1 - \left(1 + \xi \left(\mu_C + \sigma_C \frac{\phi(\bar{C}_{\min}) - \phi(\bar{C}_{\max})}{\Phi(\bar{C}_{\max}) - \Phi(\bar{C}_{\min})}\right) \left(\frac{1}{\mu_D} + \frac{\text{Var}(D)}{\mu_D^3}\right) T_d\right)^{-\frac{1}{\xi}} \quad (27)$$

for the case of  $\xi \neq 0$ . When  $\xi = 0$ , similar to Eq. (15) it can be derived by plugging Eq. (25) into Eq. (2).

### C. Uncertainty and Confidence Assessment

Having presented two approaches for modeling task assignment reliability in XEC systems, a basic approach using uniform distributions when only minimal information is available, and a more refined approach using truncated normal distributions when historical data is available, we now turn our attention to evaluating the effectiveness of these estimations in providing higher confidence. This evaluation is crucial for understanding the uncertainty associated with our reliability metric and assessing the confidence we can place in our estimates.

To quantify the uncertainty in our estimates, we employ the concept of entropy from information theory. Entropy provides

a measure of the average amount of information contained in a random variable and serves as an excellent proxy for the uncertainty in our estimates. Lower entropy indicates less uncertainty and, consequently, higher confidence in our reliability metric.

We begin with a fundamental result from information theory [43]:

For a continuous random variable  $X$  and a differentiable, invertible function  $g$ , the entropy of  $Y = g(X)$  is given by:

$$H(Y) = H(X) + \mathbb{E}_X[\log |g'(X)|] \quad (28)$$

where  $g'(X)$  is the derivative of  $g$  with respect to  $X$ , and  $\mathbb{E}_X$  denotes expectation with respect to  $X$ .

This theorem implies that the entropy of a transformed random variable  $Y = g(X)$  is influenced by the entropy of  $X$  and the average of the logarithm of the absolute derivative of the transformation, that is  $H(Y) = H(g(X)) \leq H(X)$ . It is crucial in our subsequent analysis of how transformations affect uncertainty.

Now, let us consider the entropy of our uniform and truncated normal distributions:

The entropy of a uniform distribution  $X_U \sim U(a, b)$  is given by:

$$H(X_U) = \log(b - a) \quad (29)$$

The uniform distribution maximizes entropy among all continuous distributions supported on the interval  $[a, b]$ , representing the greatest uncertainty given only the interval.

The entropy of a truncated normal distribution  $X_{TN} \sim \text{TNORM}(\mu, \sigma^2, a, b)$  is given by:

$$H(X_{TN}) = \frac{1}{2} \log(2\pi e \sigma^2) + \log\left(\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right) - \frac{\left[(b-\mu)\phi\left(\frac{b-\mu}{\sigma}\right) - (a-\mu)\phi\left(\frac{a-\mu}{\sigma}\right)\right]}{\sigma \left[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right]} \quad (30)$$

where  $\phi(\cdot)$  and  $\Phi(\cdot)$  are the probability density function and cumulative distribution function of the standard normal distribution, respectively.

With these entropies defined, we can now state our main lemma regarding the relative uncertainties of our approaches:

**Lemma 1.** *Let  $X_{TN} \sim \text{TNORM}(\mu, \sigma^2, a, b)$ ,  $\sigma < \infty$  and  $X_U \sim U(a, b)$  over the same interval  $[a, b]$ . Then, the entropy of  $X_{TN}$  is upper bounded by the entropy of  $X_U$ :*

$$H(X_{TN}) < H(X_U) \quad (31)$$

*Proof.* Since the uniform distribution maximizes entropy over  $[a, b]$ , any other distribution on the same interval has entropy less than or equal to  $H(X_U)$ . Therefore,  $H(X_{TN}) < H(X_U)$   $\square$ .

This result demonstrates that our approach with historical data, using truncated normal distributions, provides estimates with lower entropy (and thus lower uncertainty) compared to the MI approach using uniform distributions.

Building on this result, we can extend our analysis to the ratio of random variables, which is central to our reliability metric:

**Lemma 2.** For independent and positive random variables  $X_{TN}, Y_{TN}$  following truncated normal distributions over  $[a, b]$ , and  $X_U, Y_U$  following uniform distributions over the same interval, the entropy of their ratio satisfies:

$$H\left(\frac{X_{TN}}{Y_{TN}}\right) \leq H\left(\frac{X_U}{Y_U}\right) \quad (32)$$

*Proof.* Let  $X_{TN}$  and  $Y_{TN}$  be independent positive random variables following truncated normal distributions over the interval  $[a, b]$ , and let  $X_U$  and  $Y_U$  be independent positive random variables following uniform distributions over the same interval  $[a, b]$ .

First, we know from Lemma 1 that:

$$H(X_{TN}) < H(X_U), \quad H(Y_{TN}) < H(Y_U) \quad (33)$$

Next, Consider the random variables  $Z_{TN} = \frac{X_{TN}}{Y_{TN}}$  and  $Z_U = \frac{X_U}{Y_U}$ . Since  $X_{TN}$  and  $Y_{TN}$  are independent, the joint entropy is:

$$H(X_{TN}, Y_{TN}) = H(X_{TN}) + H(Y_{TN}) \quad (34)$$

Similarly:

$$H(X_U, Y_U) = H(X_U) + H(Y_U) \quad (35)$$

From Theorem 1, For a transformation  $Z = \frac{X}{Y}$ , the entropy of  $Z$  satisfies:

$$H(Z) \leq H(X, Y) \quad (36)$$

because the entropy may decrease under a deterministic transformation.

Therefore:

$$H\left(\frac{X_{TN}}{Y_{TN}}\right) \leq H(X_{TN}, Y_{TN}) = H(X_{TN}) + H(Y_{TN}) \quad (37)$$

$$H\left(\frac{X_U}{Y_U}\right) \leq H(X_U, Y_U) = H(X_U) + H(Y_U) \quad (38)$$

Since  $H(X_{TN}) + H(Y_{TN}) \leq H(X_U) + H(Y_U)$ , it follows that:

$$H\left(\frac{X_{TN}}{Y_{TN}}\right) \leq H\left(\frac{X_U}{Y_U}\right) \quad \square \quad (39)$$

These theoretical results have significant practical implications for our reliability estimation in XEC systems: (1) Higher certainty in estimates: By leveraging truncated normal distributions where data is available, we consistently achieve reliability estimates with higher certainty (lower entropy) compared to the MI scenario that uses uniform distributions, (2) Improved confidence with more information: As we gather more information to refine our parameter estimates for the truncated normal distributions, we can expect further reductions in uncertainty, leading to even more confident reliability assessments.

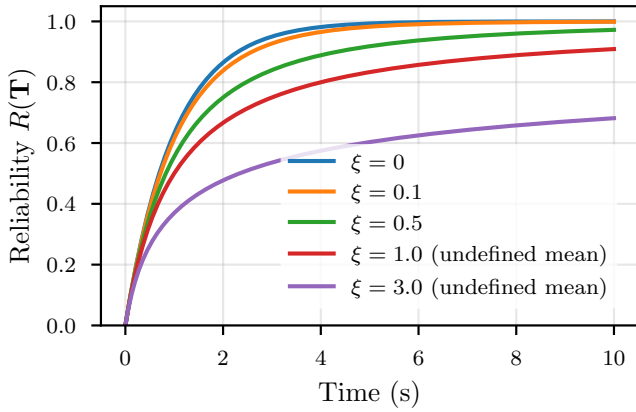
In practical terms, this improvement in confidence, achieved by leveraging historical data when available, allows for more reliable decision-making in task assignment and resource management within XEC systems. The orchestrator can make more informed choices about which workers to assign tasks to, based on more accurate and confident reliability estimates.

This, in turn, leads to improved system performance, reduced task failures, and better overall quality of service in XEC environments. To illustrate, consider a worker whose reliability is estimated as  $R(T_d) = 0.9$ . Under the MI scenario (high entropy), there is substantial uncertainty around this estimate, meaning the true reliability could deviate significantly. Under the historical data scenario (lower entropy), the same value of  $R(T_d) = 0.9$  carries more confidence, enabling the orchestrator to make more decisive task assignment decisions.

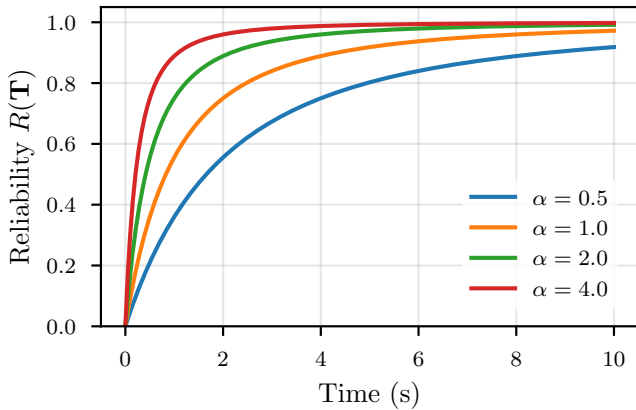
## V. SIMULATION AND EXPERIMENT RESULTS

In this section, we present validation of our reliability models through Monte Carlo simulation. The Monte Carlo approach allows us to handle the complex probabilistic nature of our models by generating multiple independent trials. For each configuration, we conduct 50-1000 trials depending on the specific experiment, providing statistical robustness while maintaining computational efficiency. For each experiment, computational capacities and task demands are sampled from their respective distributions, that is, uniform distributions for MI scenarios and truncated normal distributions when data is available scenarios. The validation strategy compares analytical predictions against simulation results across various parameter ranges and system conditions. For task assignment, we examine both scenarios through randomized sampling from uniform and truncated normal distributions respectively. The analytical versus simulation comparison serves two purposes: validating the mathematical tractability of our derived expressions and assessing their practical applicability in real-world scenarios. This dual validation approach helps identify any gaps between theoretical predictions and practical system behavior, especially in edge cases where simplifying assumptions might impact model accuracy.

In Figure 2, we can observe the fundamental characteristics of the Generalized Pareto Distribution (GPD) and its implications for reliability modeling in XEC systems. In Figure 2(a), we demonstrate how the shape parameter  $\xi$  affects the reliability function when  $\alpha$  is fixed at 1.0. It can be noticed that for lower values of  $\xi$  (0 and 0.1), the reliability function rapidly approaches 1, indicating a more predictable behavior with less variance in task completion times. Indeed, when  $\xi = 0$ , the distribution reduces to an exponential distribution, showing the classical memoryless property. As  $\xi$  increases (0.5 and above), the reliability function exhibits a heavier tail, and for  $\xi \geq 1$ , where the mean becomes undefined, the asymptotic reliability significantly decreases, suggesting a non-negligible probability of extremely long task completion times. Subsequently, Figure 2(b) illustrates the effect of the scale parameter  $\alpha$  on the reliability function with a fixed shape parameter  $\xi = 0.5$ . We can observe that higher values of  $\alpha$  lead to faster convergence to the asymptotic reliability value. In fact, as  $\alpha$  increases from 0.5 to 4.0, the rate at which tasks are completed successfully improves substantially, particularly in the early time period (0-2 seconds). This behavior aligns with our theoretical analysis, as  $\alpha$  represents the execution rate of the XED, where higher values indicate more efficient task processing capabilities. The most significant improvement in reliability is observed when



(a) Effect of Shape Parameter  $\xi$  ( $\alpha = 1.0$ )



(b) Effect of Scale Parameter  $\alpha$  ( $\xi = 0.5$ )

Fig. 2. Generalized Pareto Distribution (GPD) reliability function with (a) varying shape  $\xi$ , fixed  $\alpha = 1.0$  (b) varying scale  $\alpha$ , fixed  $\xi = 0.5$

increasing  $\alpha$  from 0.5 to 2.0, while further increases (from 2.0 to 4.0) show diminishing returns in terms of reliability enhancement.

Afterwards, we validate our reliability analytical model with MI. For the results shown in Figure 3, we used practical parameter values derived from typical edge device capabilities: computational capacity bounds  $C_{\min} = 1.2$  GHz and  $C_{\max} = 3.6$  GHz, and task demand bounds  $D_{\min} = 1.0$  GHz and  $D_{\max} = 3.0$  GHz. The simulation results were averaged over 50 independent trials to ensure statistical significance. Figure 3 compares the analytical and simulation results for the MI-based reliability model under different shape parameters  $\xi$ . The dashed lines represent our analytical derivations, while the triangular markers show the mean simulation results, with shaded regions indicating one standard deviation confidence intervals. We can observe the alignment between the analytical and simulation results across all tested  $\xi$  values. For  $\xi = 0$ , which corresponds to exponential task completion behavior, both analytical and simulation results show rapid convergence to near-perfect reliability within 4 seconds. When  $\xi$  increases to 0.5, the reliability curve exhibits a more gradual rise, reaching about 95% reliability at 8 seconds. Most notably, with  $\xi = 1.0$ , where the mean task completion time becomes

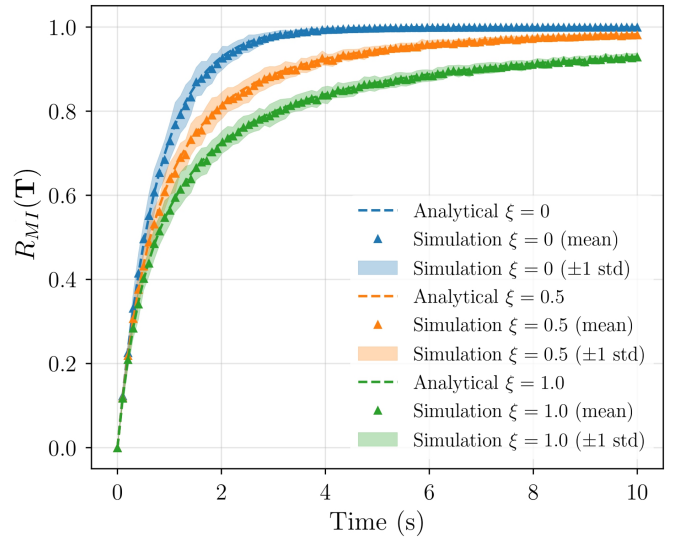
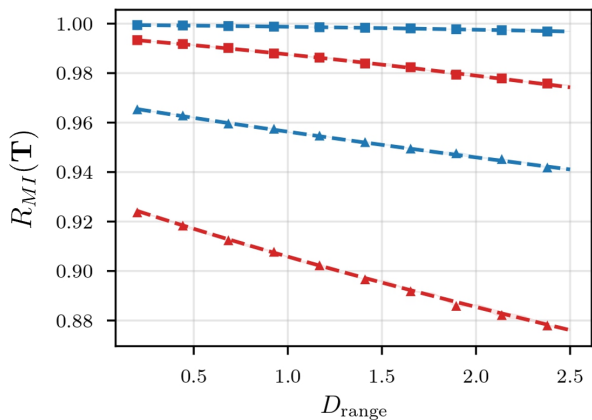


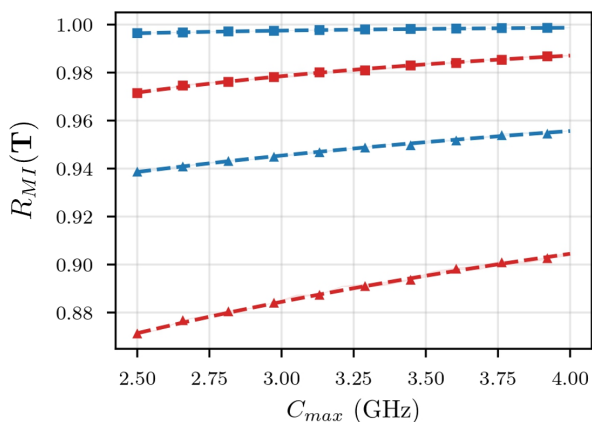
Fig. 3. Simulation vs analytical reliability function with MI for varying  $\xi \in \{0, 0.5, 1.0\}$

undefined theoretically, the system achieves only about 90% reliability even after 10 seconds, demonstrating the heavy-tailed nature of the distribution. The narrow confidence intervals (shaded regions) in our simulation results, particularly for lower  $\xi$  values, validate the robustness of our analytical model. Indeed, the slight widening of confidence intervals as  $\xi$  increases aligns with our theoretical expectations, as higher shape parameters introduce more variability in task completion times. These results confirm that our MI-based analytical framework accurately captures the reliability characteristics of XEC systems even with minimal information about worker capabilities and task demands. Notably, the comparison across  $\xi$  values in Figures 3 and 9 also serves as a built-in comparative analysis: the  $\xi = 0$  case corresponds to the exponential model commonly assumed in prior work. As can be observed, relying solely on the exponential assumption would overestimate reliability for workloads that exhibit heavier-tailed behavior ( $\xi > 0$ ), underscoring the value of the more general GPD formulation.

We then examine the impacts of demand variability and computational capacity on reliability in Figure 4. The simulations were conducted with base parameters  $C_{\min} = 2.0$  GHz, initial  $D_{\min} = 1.5$  GHz, and reliability was evaluated at two time points ( $T_d = 5$  and  $T_d = 10$  seconds) with two shape parameters ( $\xi = 0.2$  and  $\xi = 0.8$ ). Figure 4(a) reveals how increasing task demand variability affects system reliability. As the demand range ( $D_{\text{range}}$ ) widens from 0.2 to 2.5 GHz, we observe a consistent decrease in reliability across all configurations. This degradation is more pronounced for shorter deadlines ( $T_d = 5$ ) and larger shape parameters ( $\xi = 0.8$ ), where reliability drops from approximately 92% to 88%. The simulation results (markers) closely track our analytical predictions (dashed lines), validating our theoretical framework. The impact of maximum computational capacity is demonstrated in Figure 4(b), where  $C_{\max}$  varies from 2.5 to 4.0 GHz. In contrast to demand variability, increasing



(a) Effect of Demand Range



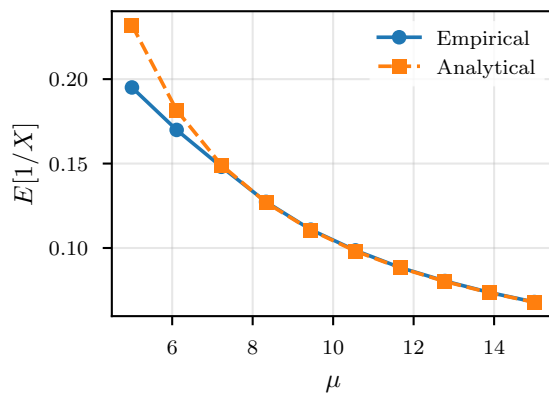
(b) Effect of Maximum Capacity

- Analytical ( $T = 5, \xi = 0.2$ )      --- Analytical ( $T = 10, \xi = 0.2$ )
- Simulation ( $T = 5, \xi = 0.2$ )      ■ Simulation ( $T = 10, \xi = 0.2$ )
- Analytical ( $T = 5, \xi = 0.8$ )      --- Analytical ( $T = 10, \xi = 0.8$ )
- ▲ Simulation ( $T = 5, \xi = 0.8$ )      ▲ Simulation ( $T = 10, \xi = 0.8$ )

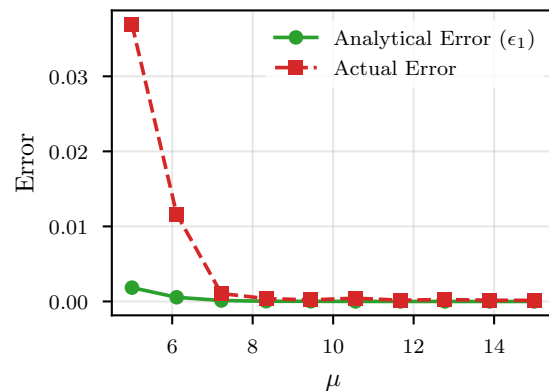
Fig. 4. Simulation vs analytical with MI: Effects of (a) varying  $D_{\text{range}}$ , fixed  $C_{\text{min}} = 2.0$  GHz (b) varying  $C_{\text{max}}$ , fixed  $D$  bounds

$C_{\text{max}}$  improves system reliability. This improvement is particularly evident for configurations with higher  $\xi$  values, where reliability increases by approximately 4 percentage points across the tested range. For longer deadlines ( $T_d = 10$ ), the system maintains near-perfect reliability regardless of capacity increases, suggesting that extended deadlines can compensate for computational limitations. The alignment between analytical predictions and simulation results, evidenced by the overlapping markers and dashed lines, further validates our MI-based modeling approach.

To validate our analytical approximations for the model with data, we conduct error analysis of the first and second moments, as well as their ratios. The simulations utilize parameters  $\sigma = 2.0$  and truncation bounds  $[3, 20]$  for varying mean values  $\mu$  ranging from 5 to 15. Figure 5 examines the accuracy of our first moment approximation  $E[1/X]$ . The analytical predictions closely match the empirical results as shown in Figure 5(a), with both curves exhibiting consistent monotonic decrease as  $\mu$  increases. The error analysis in



(a) Simulation vs Analytical Approximation of  $E[1/X]$

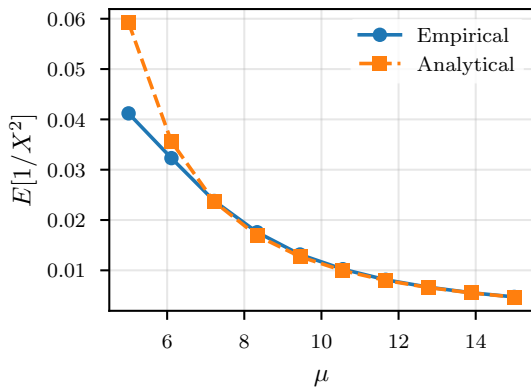


(b) Error Analysis of  $E[1/X]$  Approximation

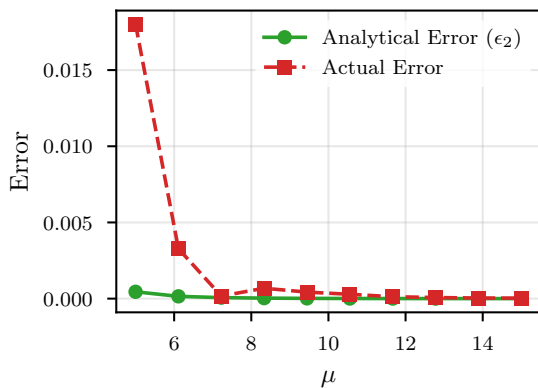
Fig. 5. Approximation error analysis of  $E[1/X]$ : (a) simulation vs analytical (b) actual error vs analytical error bound. Varying  $\mu$ , fixed  $\sigma = 2.0$ , bounds  $[3, 20]$

Figure 5(b) reveals that the actual error remains below 0.5% for  $\mu > 8$ , with the analytical error bound  $\epsilon_1$  providing a reliable upper bound that becomes increasingly tight at higher  $\mu$  values. Similarly, Figure 6 presents the analysis for the second moment approximation  $E[1/X^2]$ . Figure 6(a) demonstrates strong consistency between analytical and empirical results, though with slightly larger discrepancies at lower  $\mu$  values compared to the first moment case. The error analysis in Figure 6(b) shows that the actual error follows a similar pattern but with marginally higher magnitudes, remaining below 0.2% for  $\mu > 8$ . The analytical error bound  $\epsilon_2$  again provides a consistent upper bound on the approximation error. For the ratio analysis shown in Figure 7, we examine the relationships between two truncated normal random variables with  $\mu_2 = 15.0$ ,  $\sigma_2 = 3.0$ , and truncation bounds  $[1, 30]$  for the denominator. Figure 7(a) demonstrates the approximation accuracy for  $E[X_1/X_2]$ , showing a linear increase with  $\mu_1$  and remarkable agreement between analytical and empirical results. The second-order ratio analysis  $E[X_1^2/X_2^2]$  in Figure 7(b) exhibits similar accuracy but with a quadratic growth pattern, validating our theoretical framework for more complex ratio calculations.

Moving forward, Figure 8 illustrates the effectiveness of Maximum Likelihood Estimation (MLE) in characterizing



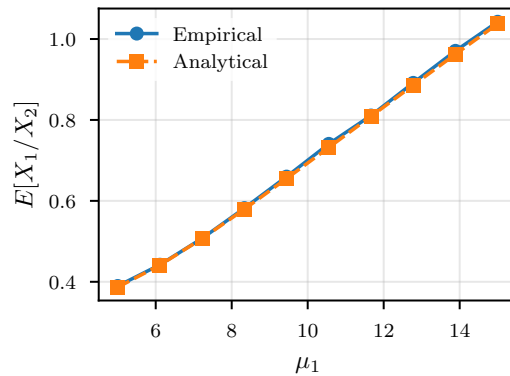
(a) Simulation vs Analytical Approximation of  $E[1/X^2]$



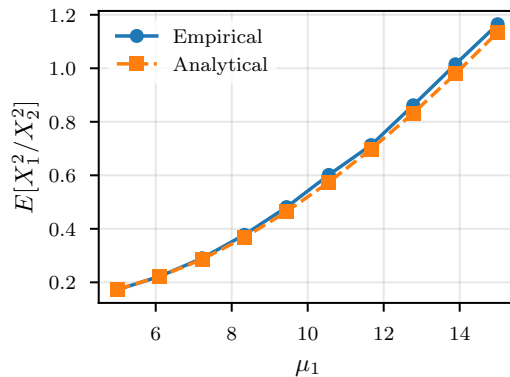
(b) Error Analysis of  $E[1/X^2]$  Approximation

Fig. 6. Approximation error analysis of  $E[1/X^2]$ : (a) simulation vs analytical (b) actual error vs analytical error bound. Varying  $\mu$ , fixed  $\sigma = 2.0$ , bounds  $[3, 20]$

the true distribution parameters from limited samples. The simulations were conducted with true parameters  $\mu = 2.2$  and  $\sigma = 0.5$ , bounded within  $[1.0, 4.0]$ . The figure contrasts the uniform distribution assumption used in MI (gray shaded area) against the true truncated normal distribution (red line) and MLE fits with increasing sample sizes. With just 3 samples (blue line), the MLE produces a narrower, more peaked distribution compared to the true distribution, reflecting the limited information available. As the sample size increases to 10 samples (green line), the estimated distribution begins to better capture the shape of the true distribution, though still exhibiting some bias in both location and scale parameters. At 20 samples (purple line), the MLE achieves remarkable conformity with the true distribution, demonstrating the rapid convergence of our parameter estimation approach. The scattered markers at the bottom of the plot represent the actual samples used for each estimation, highlighting how the MLE effectively leverages even sparse data points to approximate the underlying distribution. This progression from uniform (MI) to increasingly accurate truncated normal distributions underscores the value of incorporating additional information into the reliability modeling. In practice, this illustrates the natural lifecycle of a worker in the XEC system: upon joining, the worker is assessed using the MI model with its declared



(a) Simulation vs Analytical Approximation of  $E[X_1/X_2]$



(b) Simulation vs Analytical Approximation of  $E[X_1^2/X_2^2]$

Fig. 7. Approximation error analysis of (a)  $E[X_1/X_2]$  and (b)  $E[X_1^2/X_2^2]$ . Varying  $\mu_1$ , fixed  $\mu_2 = 15.0$ ,  $\sigma_2 = 3.0$ , bounds  $[1, 30]$

bounds, and as task interactions accumulate, the orchestrator transitions to the historical data model with progressively refined reliability estimates.

To evaluate our reliability model with data availability, we conducted simulation with the following: computational capacity with mean  $\mu_C = 2.4$  GHz and standard deviation  $\sigma_C = 0.6$  GHz bounded within  $[1.2, 3.6]$  GHz, and task demand with mean  $\mu_D = 2.0$  GHz and standard deviation  $\sigma_D = 0.5$  GHz bounded within  $[1.0, 3.0]$  GHz. The simulation results were averaged over 50 independent trials to ensure statistical

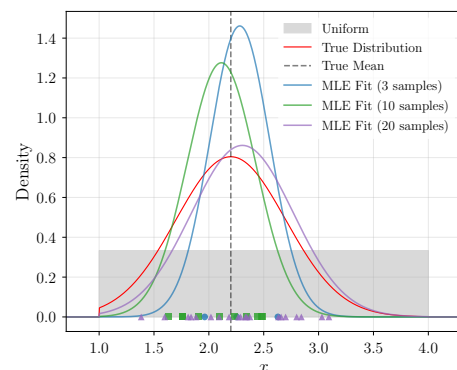


Fig. 8. MLE estimation with progressing number of samples. True parameters  $\mu = 2.2$ ,  $\sigma = 0.5$ , bounds  $[1.0, 4.0]$

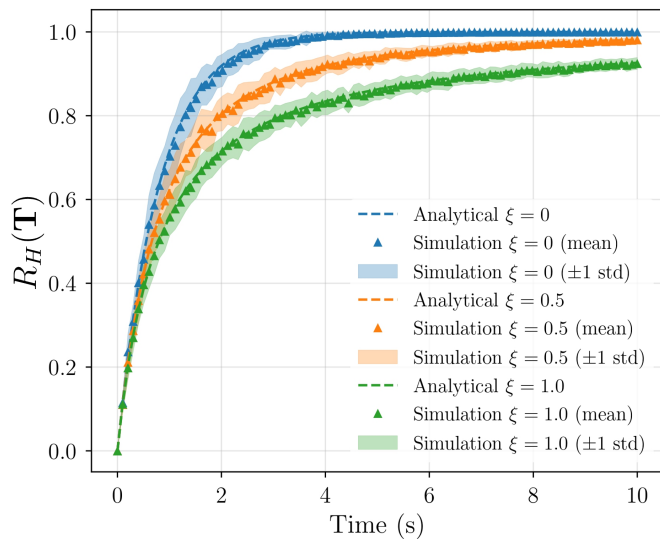
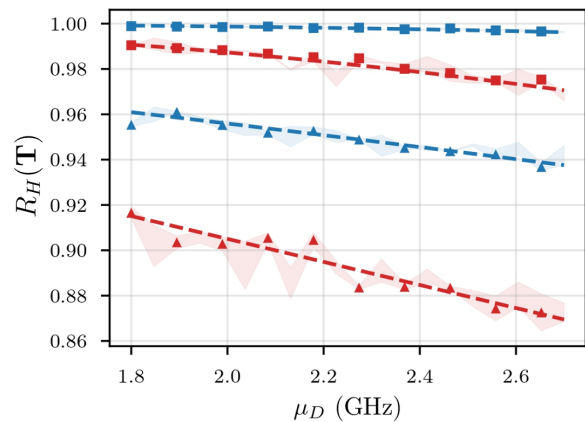


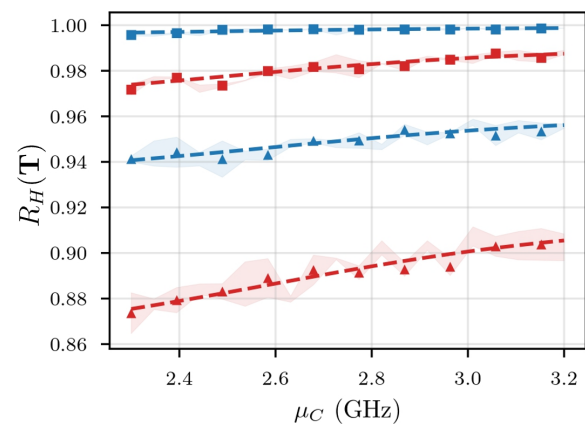
Fig. 9. Simulation vs analytical reliability function with historical data for varying  $\xi \in \{0, 0.5, 1.0\}$

robustness. Figure 9 compares the analytical predictions with simulation results for different shape parameters  $\xi$ . For the exponential case ( $\xi = 0$ ), both analytical and simulation results show rapid convergence to near-perfect reliability, reaching approximately 99% within 4 seconds. The model with  $\xi = 0.5$  exhibits a more gradual increase, achieving about 95% reliability at 8 seconds. When  $\xi = 1.0$ , where the mean completion time becomes theoretically undefined, the system attains roughly 90% reliability after 10 seconds, demonstrating the heavy-tailed nature of the distribution. The narrow confidence intervals (shaded regions) across all  $\xi$  values, particularly in the early time periods ( $t \leq 4s$ ), validate the stability of our data-based analytical framework. The slight widening of confidence intervals at higher  $\xi$  values aligns with theoretical expectations, as larger shape parameters introduce increased variability in task completion times. Notably, the analytical predictions (dashed lines) maintain strong correspondence with simulation means (triangular markers) throughout the entire time range, demonstrating the accuracy of our truncated normal approximation approach.

Figure 10 examines how mean demand and computational capacity affect the reliability. The simulations use computational bounds  $[2.0, 3.5]$  GHz with  $\sigma_C = 0.3$  GHz, and demand bounds  $[1.5, 3.0]$  GHz with  $\sigma_D = 0.25$  GHz. Results are presented for two time deadlines ( $T_d = 5, 10$  seconds) and shape parameters ( $\xi = 0.2, 0.8$ ). Figure 10(a) reveals that increasing the mean demand  $\mu_D$  from 1.8 to 2.7 GHz reduces reliability across all configurations. For  $T_d = 10$  and  $\xi = 0.2$ , reliability remains steady at approximately 99%, while for  $T_d = 5$  and  $\xi = 0.8$ , it drops from 92% to 87%. The analytical predictions (dashed lines) match the simulation results (markers), with narrow confidence intervals indicating stable predictions. Conversely, Figure 10(b) shows that increasing the mean capacity  $\mu_C$  from 2.3 to 3.2 GHz improves reliability. The improvement is more pronounced for shorter deadlines and higher shape parameters, where reliability rises from 87%



(a) Effect of Demand Mean



(b) Effect of Capacity Mean

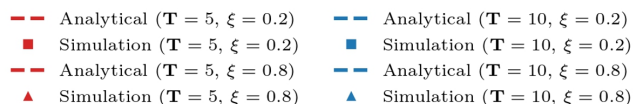


Fig. 10. Simulation vs analytical with historical data: Effects of (a) varying  $\mu_D$  (b) varying  $\mu_C$ , for  $T_d \in \{5, 10\}$ s and  $\xi \in \{0.2, 0.8\}$

to 90% ( $T_d = 5, \xi = 0.8$ ). For longer deadlines ( $T_d = 10$ ), the system maintains high reliability regardless of capacity increases. The conformity between analytical predictions and simulations, coupled with the consistent confidence intervals, validates the truncated normal approximation approach for reliability estimation.

#### A. Experimental Validation with GEMM Workloads

To validate the proposed framework beyond synthetic simulations, we developed a software experiment using Docker containers to emulate XED environments. In this setup, the computational resources of the container are varied to represent different device states, and the task demand is varied to represent different workload requirements. Without loss of generality, we use General Matrix Multiplication (GEMM) as the benchmark task, as it abstracts many compute-intensive workloads commonly found at the edge, such as neural network inference, image processing, and signal processing.

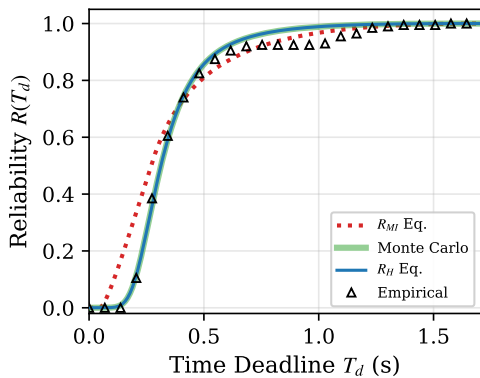


Fig. 11. Experimental validation: Reliability comparison using GEMM workloads in Docker containers. Four curves are shown: MI analytical model, Historical analytical model, Monte Carlo simulation, and empirical reliability from 200 GEMM tasks.

The experiment uses 200 tasks with variable capacity  $C$  controlled via thread allocation (1–6 threads, calibrated at 71.7–446.8 GFLOPS), with non-uniform selection reflecting realistic device operation patterns where lower-capacity states are more frequent. Variable demand  $D$  is achieved through matrix sizes sampled from a truncated normal distribution over [20, 120] GFLOPs.

Figure 11 compares four reliability curves: (1) the MI analytical model assuming uniform distributions over declared bounds, (2) the Historical analytical model using MLE-fitted distributions from observed data, (3) a Monte Carlo simulation sampling 500k  $(C, D)$  pairs from the fitted distributions and computing  $T = D/C$ , and (4) the empirical reliability from the actual 200 GEMM execution times. The results show strong alignment among the data-driven curves, validating our analytical framework with real workloads. The MI model, as expected, overestimates reliability due to its uniform assumption over the full declared bounds.

## VI. DISCUSSION

Quantifying the computational reliability provides a method for deploying resource-intensive applications, such as Federated Learning (FL) and distributed Large Language Models (LLMs) inference, in XEC environments. By translating the uncertain performance of consumer-owned devices into a probabilistic metric, this work addresses operational challenges in guaranteeing service dependability. Without a formal means to assess worker reliability, orchestrating multi-node computations would depend on heuristics or conservative resource allocations.

For FL, the models presented here offer a quantitative approach to participant selection. Operationally, an orchestrator can set a minimum reliability threshold for inclusion, using the metric  $R(T_d)$  to form a cohort of devices with a high probability of completing training tasks within a synchronous round. This mitigates the effects of stragglers by preemptively filtering out unpredictable devices, which can lead to improved convergence times and more efficient resource utilization. For distributed LLM inference, which often relies on pipeline

and tensor parallelism [44], the framework allows for an a priori assessment of end-to-end system viability. The orchestrator can compose a viable processing chain by ensuring the product of individual node reliabilities meets a system-wide target. This enables the construction of dependable inference pipelines by identifying and excluding nodes with low reliability scores, a function necessary for supporting interactive applications such as chatbots that are sensitive to stalls or delays.

More broadly, the reliability metric  $R(T_d)$  can be operationalized by an orchestrator in several ways: (1) *worker ranking*, where available workers are sorted by their reliability for a given deadline and the highest-ranked worker is selected; (2) *threshold-based filtering*, where only workers exceeding a minimum reliability threshold (e.g.,  $R(T_d) \geq 0.9$ ) are considered for task assignment; and (3) *confidence-aware scheduling*, where the reduced uncertainty from historical data (as shown in Section IV-C) allows the orchestrator to make more decisive assignments.

It is worth noting that the MI bounds provided by workers serve as guarantees: a worker cannot promise a single deterministic value for its resources, but rather a window within which it will operate. If a worker is unsure about its capabilities, the basic choice is to set the window to the absolute minimum and maximum values for its resources. If a worker provides a window and subsequently violates it, this breaks the initial contract, which could harm the worker’s reputation. Reputation-based mechanisms for handling such violations represent a promising direction for future work.

While the proposed framework provides a principled approach to reliability estimation, several limitations should be acknowledged. First, non-stationary behavior of XEDs over long time horizons may cause the historical data model parameters to drift; this is addressed via periodic re-estimation of MLE parameters using a sliding-window approach. Second, successive capacity observations may exhibit correlation under rapid workload changes; this is mitigated by the temporal separation between task executions as noted in Section IV-B. Third, the current work focuses on single-worker reliability; multi-worker composition, redundancy strategies, and load balancing are important topics that naturally build on this per-worker metric and represent promising directions for future work.

## VII. CONCLUSION

This paper addressed the challenge of quantifying computational reliability for task assignment in Extreme Edge Computing (XEC) systems, which are characterized by unpredictable and heterogeneous devices operating with incomplete information. We proposed a mathematical framework that models reliability using the Generalized Pareto Distribution (GPD) to accommodate diverse task completion behaviors. The framework offers two models: a baseline derived from operational bounds for scenarios with minimal information, and a data-driven model that leverages historical observations via Maximum Likelihood Estimation (MLE) to refine estimates. We proved that this data-driven approach systematically

reduces estimation uncertainty. The resulting models provide an analytically tractable method for service orchestrators to make risk-aware task assignment decisions. By translating the stochastic nature of Extreme Edge Devices (XEDs) into a quantifiable reliability metric, this work establishes a foundational component for building dependable XEC systems. This is a prerequisite for enabling advanced applications and for realizing future paradigms like the Computing Power Network (CPN), where the dependability of each node must be assessed.

#### ACKNOWLEDGMENT

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN-2025-05001.

#### REFERENCES

- [1] D. Wen et al., "Task-oriented sensing, computation, and communication integration for multi-device edge ai," *IEEE Transactions on Wireless Communications*, vol. 23, no. 3, pp. 2486–2502, 2023.
- [2] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016. DOI: 10.1109/JIOT.2016.2579198
- [3] K. Cao, Y. Liu, G. Meng, and Q. Sun, "An overview on edge computing research," *IEEE access*, vol. 8, pp. 85 714–85 728, 2020.
- [4] J. Liu et al., "Computing power network: A testbed and applications with edge intelligence," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, IEEE, 2022, pp. 1–2.
- [5] J. Portilla, G. Mujica, J.-S. Lee, and T. Riesgo, "The extreme edge at the bottom of the internet of things: A review," *IEEE Sensors Journal*, vol. 19, no. 9, pp. 3179–3190, 2019. DOI: 10.1109/JSEN.2019.2891911
- [6] S. B. Azmy, R. F. El-Khatib, N. Zorba, and H. S. Hassanein, "Extreme edge computing challenges on the edge-cloud continuum," in *2024 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2024, pp. 99–100. DOI: 10.1109/CCECE59415.2024.10667328
- [7] S. B. Azmy, N. Zorba, and H. S. Hassanein, "Incentive-vacation queueing in extreme edge computing: An analytical reward-based framework," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 2183–2195, 2024. DOI: 10.1109/OJCOMS.2024.3383046
- [8] E. Covi et al., "Adaptive extreme edge computing for wearable devices," *Frontiers in Neuroscience*, vol. 15, p. 611 300, 2021.
- [9] O. Naserallah, S. B. Azmy, N. Zorba, and H. S. Hassanein, "Impact of users' mobility on the quality of edge sensing systems," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 2022, pp. 4196–4201. DOI: 10.1109/GLOBECOM48099.2022.10000854
- [10] M. S. Allahham, A. Mohamed, A. Erbad, and H. Hassanein, "On the modeling of reliability in extreme edge computing systems," in *2022 5th International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, 2022, pp. 1–6. DOI: 10.1109/ICCSPA55860.2022.10019108
- [11] Y.-F. Li and R. Peng, "Service reliability modeling of distributed computing systems with virus epidemics," *Applied Mathematical Modelling*, vol. 39, no. 18, pp. 5681–5692, 2015.
- [12] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE communications surveys & tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.
- [13] W. Yu et al., "A survey on the edge computing for the internet of things," *IEEE Access*, vol. 6, pp. 6900–6919, 2018. DOI: 10.1109/ACCESS.2017.2778504
- [14] R. Tourani et al., "Democratizing the edge: A pervasive edge computing framework," *arXiv preprint arXiv:2007.00641*, 2020.
- [15] A. A. Moustafa, S. A. Elsayed, and H. S. Hassanein, "Community-oriented resource allocation at the extreme edge," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 2022, pp. 5583–5588. DOI: 10.1109/GLOBECOM48099.2022.10001198
- [16] R. F. El Khatib, S. A. Elsayed, N. Zorba, and H. S. Hassanein, "Optimal proactive resource allocation at the extreme edge," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 5657–5662. DOI: 10.1109/ICC45855.2022.9838897
- [17] M. De'bas, S. A. Elsayed, and H. S. Hassanein, "Multitiered worker-oriented resource allocation at the extreme edge," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 2022, pp. 5674–5679. DOI: 10.1109/GLOBECOM48099.2022.10000855
- [18] I. M. Amer and S. Sorour, "Cost-based compute cluster formation in edge computing," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 1611–1616. DOI: 10.1109/ICC45855.2022.9838830
- [19] R. Kain and S. Sorour, "Worker resource characterization under dynamic usage in multi-access edge computing," in *2022 International Wireless Communications and Mobile Computing (IWCMC)*, 2022, pp. 1070–1075. DOI: 10.1109/IWCMC55113.2022.9824299
- [20] R. Kain, S. A. Elsayed, Y. Chen, and H. S. Hassanein, "Rump: Resource usage multi-step prediction in extreme edge computing," *Computer Communications*, vol. 210, pp. 45–57, 2023, ISSN: 0140-3664. DOI: <https://doi.org/10.1016/j.comcom.2023.07.029> [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S014036642300261X>
- [21] R. Kain, S. A. Elsayed, Y. Chen, and H. S. Hassanein, "Multi-step prediction of worker resource usage at the extreme edge," in *Proceedings of the 25th International ACM Conference on Modeling Analysis and Simulation of Wireless and Mobile Systems*, ser. MSWiM '22, Montreal, Quebec, Canada: Association for Computing Machinery, 2022, pp. 25–32, ISBN: 9781450394826. DOI: 10.1145/3551659.3559051 [Online]. Available: <https://doi.org/10.1145/3551659.3559051>
- [22] A. E. Zonouz, L. Xing, V. M. Vokkarane, and Y. L. Sun, "Reliability-oriented single-path routing protocols in wireless sensor networks," *IEEE Sensors Journal*, vol. 14, no. 11, pp. 4059–4068, 2014. DOI: 10.1109/JSEN.2014.2332296
- [23] N. Jiang, Y. Deng, X. Kang, and A. Nallanathan, "Random access analysis for massive iot networks under a new spatiotemporal model: A stochastic geometry approach," *IEEE Transactions on Communications*, vol. 66, no. 11, pp. 5788–5803, 2018. DOI: 10.1109/TCOMM.2018.2854275
- [24] B. Wang, M. Li, X. Jin, and C. Guo, "A reliable iot edge computing trust management mechanism for smart cities," *IEEE Access*, vol. 8, pp. 46 373–46 399, 2020. DOI: 10.1109/ACCESS.2020.2979022
- [25] M. Frustaci, P. Pace, G. Aloï, and G. Fortino, "Evaluating critical security issues of the iot world: Present and future challenges," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2483–2495, 2018. DOI: 10.1109/JIOT.2017.2767291
- [26] L. Dong, W. Wu, Q. Guo, M. N. Satpute, T. Znati, and D. Z. Du, "Reliability-aware offloading and allocation in multilevel edge computing system," *IEEE Transactions on Reliability*, vol. 70, no. 1, pp. 200–211, 2021. DOI: 10.1109/TR.2019.2909279
- [27] X. Hou et al., "Reliable computation offloading for edge-computing-enabled software-defined iot," *IEEE Internet of*

- Things Journal*, vol. 7, no. 8, pp. 7097–7111, 2020. DOI: 10.1109/JIOT.2020.2982292
- [28] M. S. Elbamby et al., “Wireless edge computing with latency and reliability guarantees,” *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1717–1737, 2019. DOI: 10.1109/JPROC.2019.2917084
- [29] C.-F. Huang, D.-H. Huang, and Y.-K. Lin, “Network reliability evaluation for a distributed network with edge computing,” *Computers & Industrial Engineering*, vol. 147, p. 106492, 2020.
- [30] J. Huang, J. Liang, and S. Ali, “A simulation-based optimization approach for reliability-aware service composition in edge computing,” *IEEE Access*, vol. 8, pp. 50355–50366, 2020. DOI: 10.1109/ACCESS.2020.2979970
- [31] J. Liu et al., “Reliability-enhanced task offloading in mobile edge computing environments,” *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 10382–10396, 2022. DOI: 10.1109/JIOT.2021.3115807
- [32] D. P. Anderson, “BOINC: A system for public-resource computing and storage,” *Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing*, pp. 4–10, 2004.
- [33] B. Javadi, D. Kondo, J.-M. Vincent, and D. P. Anderson, “Mining for statistical models of availability in large-scale distributed systems: An empirical study of SETI@home,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 20, no. 11, pp. 1650–1661, 2009.
- [34] A. Iosup, O. Sonmez, and D. Epema, “The failure trace archive: Enabling the comparison of failure measurements and models of distributed systems,” in *Journal of Parallel and Distributed Computing*, vol. 73, Elsevier, 2013, pp. 1208–1223.
- [35] M. Harchol-Balter, *Performance modeling and design of computer systems: queueing theory in action*. Cambridge University Press, 2013.
- [36] J. Yu, R. Buyya, and C. K. Tham, “Cost-based scheduling of scientific workflow applications on utility grids,” *Proceedings of the 1st IEEE International Conference on e-Science and Grid Computing*, pp. 140–147, 2005.
- [37] S. Abrishami, M. Naghibzadeh, and D. H. Epema, “Deadline-constrained workflow scheduling algorithms for infrastructure as a service clouds,” *Future Generation Computer Systems*, vol. 29, no. 1, pp. 158–169, 2013.
- [38] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, “A close examination of performance and power characteristics of 4G LTE networks,” in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services (MobiSys)*, ACM, 2012, pp. 225–238.
- [39] J. R. Hosking and J. R. Wallis, “Parameter and quantile estimation for the generalized pareto distribution,” *Technometrics*, vol. 29, no. 3, pp. 339–349, 1987. DOI: 10.1080/00401706.1987.10488243
- [40] T. Cover and J. Thomas, *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing). Wiley, 1991.
- [41] E. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003, vol. 727.
- [42] M. D. Springer, *The Algebra of Random Variables* (Wiley series in probability and Mathematical Statistics). John Wiley & Sons, 1979.
- [43] M. Thomas and A. T. Joy, *Elements of information theory*. Wiley-Interscience, 2006.
- [44] Z. Li, W. Feng, M. Guizani, and H. Yu, “Tpi-llm: Serving 70b-scale llms efficiently on low-resource edge devices,” *arXiv preprint arXiv:2410.00531*, 2024.