Scalable Dependability - 1st Workshop

October 23 and 24, 2014
PUCRS, Faculdade de Informática, Building 32, Room 207
Av. Ipiranga, 6681 - CEP 90619-900 - Porto Alegre

October 23

12:00-13:30 Lunch

13:30-14:45 1) Overview of the project; preliminary results 2) e 3)

15 min Break

15:00-16:30 4) , 5) e 6)

16:30 Discussão

October 24

9:30-10:30 7), 8)

15min Break

10:45-12:00 Technical discussion

12:00-13:30 Lunch

13:30-15:00 Next steps discussion

Abstracts

1) Scalable Dependability Project - Overview - Fernando Pedone (USI-Suíça)

2) Recovery in Parallel State-Machine Replication - Odorico Mendizabal (PUCRS/FURG), Parisa Jalili Marandi (USI), Fernando Luís Dotti (PUCRS), Fernando Pedone (USI)

   Abstract. State-machine replication is a popular approach to building fault-tolerant systems, which relies on the sequential execution of com- mands to guarantee strong consistency. Sequential execution, however, threatens performance. Recently, several proposals have suggested par- allelizing the execution model of the replicas to enhance state-machine replication?s performance. Despite their success in accomplishing high performance, the implications of these models on recovery is mostly left unaddressed. In this paper, we focus on the recovery problem in the context of Parallel State-Machine Replication. We propose two novel al- gorithms and assess them through simulation and a real implementation.

3) Analysis of Checkpointing Overhead in Parallel State-Machine Replication - Odorico Mendizabal (PUCRS/FURG), F. Pedone, F. Dotti (PUCRS)

   A well-established technique used to design fault-tolerant systems is State-Machine Replication. In part, this is ex- plained by the simplicity of the approach and its strong consistency guarantees. Recently, several proposals have suggested parallelizing the execution of state-machine repli- cas to achieve higher throughput. Concurrent execution of commands has many implications, including the proce- dure to recover replicas from failures. Conventional check- pointing techniques, for example, must be revisited in par- allelized models. In this paper, we review parallel variations of state-machine replication and discuss how checkpointing procedures apply to these models. Moreover, we evaluate the impact caused by checkpointing techniques on recovery through simulations.

4) Executando Aplicações MPI Resilientes sobre um Núcleo Dinâmico de Processos Estáveis - Edson Camargo e Elias P. Duarte Jr. - UFPR

Resumo: Diversos modelos vem sendo propostos com o objetivo de estender a norma MPI com primitivas tolerância a falhas, de modo a permitir a construção de aplicações MPI resilientes. Muitos desses modelos assumem um sistema com desempenho e disponibilidade previsíveis, onde uma falha do tipo crash corresponde a um evento permanente facilmente identificado. Neste trabalho apresentamos um modelo que assume um sistema onde os processos podem se tornar instáveis. Este é o caso de muitos sistemas reais, em particular, os clusters de alto desempenho que são compartilhados. No modelo proposto, os processos executam testes entre si e formam um núcleo de processos estáveis para a execução da aplicação. Um processo é removido do núcleo quando é considerado instável. Esta classificação pode, posteriormente, ser revertida e o consenso sobre a estabilidade de um determinado processo instável pode levar a sua re-incorporação ao núcleo. Resultados experimentais de três implementações do algoritmo ordenação paralela Hyperquicksort para ordenar 1 bilhão de números inteiros são apresentados. O modelo proposto foi implementado e resultados são comparados com a especificação de tolerância a falhas para MPI ULFM.

5) Segmentação de Overlays P2P como Suporte para Memórias Tolerantes a Intrusões - Eduardo Alchieri (UnB) + Joni Fraga (UFSC) + David Börger

Resumo: Este trabalho descreve nossa experiência no desenvolvimento de uma infraestrutura que permite a construção de memórias compartilhadas tolerantes a intrusões para sistemas de larga escala. A infraestrutura faz uso de um overlay P2P e do conceito de Replicação Máquina de Estados (RME). O conceito de segmentação é introduzido sobre o espaço de chaves do overlay para permitir o uso de algoritmos de suporte à RME. No presente trabalho descrevemos a infraestrutura proposta em sua estratificação e algoritmos. Além disso, realizamos uma análise da solução apresentada e dos custos envolvidos.

6) Lásaro Camargos (UFU) - Atomic Broadcast, an important abstraction in dependable distributed computing, is usually implemented by solving infinitely many instances of the well-known consensus problem. Some asynchronous consensus algorithms achieve the optimal latency of two (message) steps but cannot guarantee this latency even in good runs, those with timely message delivery and no crashes. This is due to *collisions*, a result of concurrent proposals. Collision-fast consensus algorithms, which decide within two steps in good runs, exist under certain conditions. Their direct application to solving atomic broadcast, however, does not guarantee delivery in two steps for all messages unless at most a single failure is tolerated. In this presentation I will show a simple way to build a fault-tolerant collision-fast Atomic Broadcast algorithm based on a variation of the consensus problem, M-Consensus, and how it may affect the future of large scale dependable computing.

7) A Formal Model for the Deferred Update Replication Technique - Andrea Corradini (Univ. Pisa) , Leila Ribeiro (UFRGS), Odorico Mendizabal (PUCRS/FURG) and Fernando Luís Dotti (PUCRS).

Database replication is a technique employed to enhance both performance and availability of database systems. The Deferred Update Replication (DUR) technique offers strong consistency (i.e. serializability) and uses an optimistic concurrency control with a lazy replication strategy relying on atomic broadcast communication. Due to its good performance, DUR has been used in the construction of several database replication protocols and is often chosen as a basic technique for several extensions considering modern environments. The correctness of the DUR technique, i.e. if DUR accepted histories are serializable, has been discussed by different authors in the literature. However, a more com- prehensive discussion involving the completeness of DUR w.r.t. serializability was lacking. As a first contribution, this paper provides an operational semantics of the DUR technique which serves as foundation to reason about DUR and its derivatives. Second, using this model the correctness of DUR w.r.t. serializabil- ity is shown. Finally, we discuss the completeness of DUR w.r.t.

serializability and show that for any serializable history there is an equivalent history accepted by DUR. Moreover, we show that transactions aborted by DUR could not be ac- cepted without changing the order of already committed transactions.

8) Model Checking the Deferred Update Replication Protocol - Odorico Mendizabal (PUCRS/FURG) e Fernando Dotti (PUCRS)

As the number of distributed applications and the volume of generated data increase, support from robust data management systems becomes even more necessary. Since these management systems possibly have to deal with heavy workloads, in this paper we analyze the deferred update replication, a successful technique to implement highly available and performing transactional databases. Although it offers a strong consistency semantics, no enough efforts have been invested to prove its properties. Due to its critical requirements, in this paper we verify the deferred update replication protocol using model checking. A model of the deferred update replication protocol and a comprehensive set of safety and liveness properties are presented. According to our investigation, all properties hold for the presented model, leading to a higher confidence in the protocol's correctness.